Diesner, J., & Carley, K. M. (2009). WYSIWII - What You See Is What It Is: Informed Approximation of Relational Data from Texts. Presentation at General Online Research (GOR) Conference, Vienna, Austria, April 2009.

WYSIWII - What You See Is What It Is: Informed Approximation of Relational Data from Texts

Jana Diesner, Kathleen M. Carley

A plethora of text data that can be collected online implicitly contain relational information, such as who says what to whom on what web site (web-version of the Lasswell formula). Potential data sources are scientific papers, legal documents, news, blogs, and emails. Relational information is often called network data. Transforming texts into networks generates concise reductions and abstractions of the original material. These network data enable us to investigate and communicate relational aspects of texts more appropriately than representations of words and documents as disjoint data points do. For cases in which we do not know what the relevant nodes and edges are, probabilistic techniques are being developed and widely used. This family of solutions involves two levels of uncertainty:

First, non-deterministic computational techniques for relation extraction (RE) only approximate network structure. This means that RE does not retrieve truth, but the most likely network given the deployed method and data. Second, validating relational data (RD) - the result of RE - by comparing them against ground truth (the true, underlying network) might be infeasible, e.g. for covert networks (price-fixing alliances, drug networks), ephemeral networks (bankrupt companies), or networks that lack an underlying real-world network (blogs). We refer to RD that cannot be validated against ground truth or that are nothing more than the gathered data as WYSIWII (What-You-See-Is-What-It-Is).

Due to the outlined uncertainties, rigorous investigations are needed so that all parties involved (developers, users, consumers of results) can understand: What variables involved in the RE process impact RD? How strong are these impacts? We address these research questions empirically and experimentally by applying different techniques for identifying, normalizing, and deduplicating relevant nodes (individuals, organizations, resources) and edges (relationships) in a corpus of several hundred research-funding descriptions. We measure the sensitivity of the resulting graphs to the applied techniques. Our findings suggest that the tested RE methods significantly impact the number and structural position of nodes and edges in networks, as well as network-level measures. We believe that our findings contribute to peoples' capability to assess the robustness and interpret the results of network data extracted from texts.

Keywords: relation extraction, natural language processing, network analysis, network validation