N15 Diesner, J., & Carley, K.M. (2013). Error propagation and robustness of relation extraction methods.

Extracting network data from natural language text data requires making several methodological choices. These choices relate to the pre-processing of text data and the identification and classification of nodes and edges. The impact of these choices on the resulting network data and analysis results is not well understood, but can have strong implications for practical applications. To address this limitation, we have conducted a series of empirical studies. We provide answers to the following questions; focusing on reference resolution and co-occurrence based link formation: First, what differences in the structure and properties of network data are due to different parameter settings for relation extraction subroutines? Second, given that relation extraction is a sequential, multi-step procedure, how do error rates of these subroutines propagate through this process? And third, how much of method-induced variations in relation extraction results can be eliminated by increasing the accuracy of these sub-routines? Our findings suggest that reference resolution on text data can change the identity and weight of 76% of the nodes and 23% of the edges in the retrieved network data, and causes major changes in common network metrics and the set of identified key entities. Minor changes in accuracy rates of reference resolution lead to comparatively huge changes in network metrics, while the set top-scoring key entities is highly robust. Co-occurrence based link formation entails a small chance of false negatives, but the rate of false positives is alarmingly high. We conclude with recommending strategies for mitigating the identified issues in practical applications.