

Diesner, J., & Carley, K. M. (2009). He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis. Proceedings of IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA), Ottawa, Canada, July 2009.

Anaphora resolution (AR) identifies the entities that pronouns refer to. Coreference resolution (CR) associates the various instances of an entity with each other. Given our data, our findings suggest that deduplicating and normalizing text data by using AR and CR impacts the literal mention, frequency, identity, and existence of about 75% of all entities in texts. Results are more moderate on the relation level: 13% of the links are modified, and 8% are removed. Performing social network analysis on the relations extracted from texts leads to findings contrary to results from corpus statistics: AR and CR cause different directions in the change of network analytical measures, AR alters these measures more strongly than CR does, and each technique identifies different sets of most crucial nodes. Bringing the results from corpus statistics and social network analysis together suggests that CR is more effective in normalizing entities, while AR is a more powerful technique for splitting up generic nodes into named entities with adjusted weights. Data changes due to AR and CR are qualitatively and quantitatively meaningful: the statistical properties of entities and relations change along with their identities. Consequently, the relational data represent the underlying social structure more truthfully. Our results can support analysts in eliminating misinterpretations of graphs distilled from texts and in selecting those nodes from social networks on which reference resolution checks should be performed.