Using grammatical clauses for social and semantic network analysis

ICA Preconference on Social and Semantic Network Analysis

Puerto Rico, 2015

Contents

Broadly described, Social and Semantic Network Analysis are methodologies for analysing a network of actors or concepts, often extracted from unstructured text. Most (text-based) social and semantic network analysis methods extract the network of actors or meanings using co-occurrence methods, for example treating actors that co-occur within a shifting window as connected or clustering words based on co-occurrence (e.g. Diesner 2013, Doerffel & Barnett 1999, Park & Leydesdorff 2013, Danowski 2012).

Although this has yielded a number of very important results, co-occurrence as a measure of association has a severe limitation, in that it only tells us if two actors or concepts are related, not how they are related. For many research questions, this is insufficient information. Especially in analyzing conflict coverage, it is quite essential to know who does something to whom: being attacked or attacking are quite different relations. More generally, in many cases we want to add a direction and a label to the edges extracted from the network.

In this paper we present a technique that we call *clause analysis*. This is a relatively simple form of grammatical analsis, where each sentence is split in clauses containing a subject and a predicate. These clauses are used to construct a combined social and semantic network: the nodes are the actors identified in subject or predicate, and a directed edge is drawn from the subject actor(s) to the predicate actor(s) with the words in the predicate forming the relation. Thus, a (trivial) example sentence "John loves Mary" would yield an edge from John to Mary labeled "love".

Technically, all sentences are preprocessed using a dependency parser such as Stanford CoreNLP, Alpino (for Dutch), or ParZU (for German). Then, actors are identified using keyword matching and coreference (anaphora) resolution (cf. Diesner & Carley, 2009). Finally, a topic model is used to reduce the dimensionality of the semantic space of the edge labels, with the words on the predicates between each actor pair in a document used as the unit of (co)-occurrence for the topic modeling.

In a sense, the problem that clause analysis tries to solve is similar to the semantic role labeling problem in computational linguistics, for which off-the-shelf tools are beginning to emerge. However, due to the variety of possible ways that especially object roles can be represented, these tools are relatively error-prone and complicated. By not attempting to differentate between the various object roles and only separating the subject (agent) from the rest, this method avoids these pitfalls.

Along with a detailed description and validation of the method, a substantive use case is presented using the international coverage of the 2009 Gaza war (Scheafer et al, 2014). Using clause analysis, the substantive research question is answered which of the actors is portrayed as the aggressor, and whether either actor is portrayed as morally bad. Earlier results have shown that newspapers in countries politically close to Israel tend to focus on Hamas' terrorism as the problem and on Israel's actions as necessary means to achive the goal of resolving the terrorist threat, while countries more politically proximate to Palestine display Israel as aggressor and focus on the suffering of Palestinian civilians.

The whole analysis can be performed within the R statistical toolkit. The first step, splitting sentences into subject/predicate clauses, is performed using AmCAT (van Atteveldt 2008) and xTas (De Rooij et al, 2012), but this service can be accessed using the API from R (or other languages). The keyword analysis, topic-modeling, and substantive analysis are all performed in R and the used scripts can be easily adopted and modified. All software and scripts are open-source and available through github, at `amcat/amcat`, `amcat/amcat-r`, `kasperwelbers/corpustools`, and `vanatteveldt/netcom15`