Towards Extracting a Semantic Graph of Entities from Unstructured Text

Ridho Reinanda, Vincent Traag, Jacky Hicks, Maarten de Rijke, Fridus Steijlen and Gerry van Klinken

Elite Network Shifts is a computational humanities project that involves collaboration between researchers in language technology, network analysis, and political science. In the Elite Network Shifts project, we aim to analyze the network of political elites as they are reported in the news media. Our corpora are news media archives compiled from different sources, including newspaper, magazines, and web articles.

An elite network is a graph with elites (political entities) as the nodes, with links between the nodes denoting the existence of any kind of association between the elites. Our corpora consist of purely textual, unstructured data; therefore a significant part of our project is devoted to automatically extracting the graph of entities from unstructured text. This requires employing an array of natural language processing and information retrieval techniques.

In the first phase of the project, we extract elite network by linking two elite nodes based on their co-occurrences within the same sentence in the corpora. We detect sentence boundaries, normalize the variations of entity names into a canonical name, and made a link between two nodes if they are mentioned together in the same documents. Analysis on the co-occurrence network extracted with this relatively simple method already yields interesting result: i.e. we are able to detect network of entities that are central in the media reports, and find clusters of elites around topics or issues.

Now, we are moving to another interesting direction: extracting semantic graph of entities. We consider two possible reasons two names can be associated together within the corpora: (i) they hold functional relationship, such as: X works for Y, A is the daughter of B, etc., or (ii) they are associated through a particular topic/issue, e.g. X and Y both are proponents of a anti-corruption legislation, X and Y were implicated in human rights violations, etc. We argue that both types of associations would be interesting for our network analysis.

The first type of associations (functional associations) involves a task that is actively being researched in the natural language processing community: semantic relation extraction. This involves automatically detecting patterns in text through a machine learning techniques. Key linguistic features in the text, such as parts-of-speech tags, named entity types, prepositions, and the structure of parse tree are extracted from the text, and later used to train a machine learning classifier.

We intend to detect the second type of entity associations (topical association) by employing information retrieval techniques: identifying aspects of an entity through topic modeling and statistical language modeling. Entity aspects are key information around an entity. In the context of the Elite Network Shifts project, this includes things like political views, involvements in cases, community roles, etc.

By extracting these two types of semantic associations between entities, we aim to enable a richer analysis centered on the semantic graph of entities.