Modelling the socio-linguistic processes using the NLP tools for syntactic parsing, Neo4j Graph database for storing and semantic and communication analysis

Benedikt Perak*[1]

[1]University of Rijeka – Croatia

Abstract

The paper deals with the application of the graph technologies for analysis of communication phenomena, primarily texts, using NLP tools, graph database and network algorithms. This approach produces tagged corpora stored in the custom model of the knowledge graph as labeled entities and property relations that can be used to explore and analyze various semantic relations within texts as well as their enriched extratextual information and correlations. More importantly for developing a humanistic type of research, this approach can be used to produce ontological model of the informational structures that can be queried across various custom based type types of entities connected to the linguistical references. This means that any linguistic level of morphosyntactic parsing can easily be related to various empirical analysis of the communication, conceptualization and framing of the social identities, interactions, institutions and cultural models.
As a case study, the paper will show the application of the graph technology for the analysis of the Croatian Parliament debates, covering sessions from the year 2003–2017. The data gathering process is published as a github project (https://github.com/rodik/Sabor). The debates of the 5th to 9th Parliamentary Assembly are downloaded as datasets of two types. In the original form, datasets are extensible but not connected in an information ontology that allows for an immediate extensive research with complex queries about the social, communicational or textual propensities. The data modelling and NLP parsing process is published in a paper by Perak and Rodik (Perak and Rodik 2018). The corpus contains data from 5 parliament Conventions, with 5599 discussion points, 895 members of the parliament, 42 political parties, 390 078 discussions. The goal of the research is to enable an extensive knowledge base with clearly structured ontology that will enable data integration, data enrichment, textual analysis and elaboration of quantitative-qualitative queries with multiple filters. We will present the process of the ontology creation using Python scripts and Neo4j database as well morphosyntactic parsing of the text using UDPipe NLP parser (Straka and Strakova 2017).

The rich social-conceptual structure of the discourse promoted by the speakers and political organizations is analyzed in terms of promoting certain perspective or a frame, using the influential words and analyzing their salience in terms of the conceptual network measures. In order to perform a Social Network Analysis (SNA) framework of the parliamentary and governmental texts, Python Igraph computational algorithms that explore the centralities

*Speaker

and detect communities of the lexical and communicational patterns (Perak and Ban Kirigin, in review) are used to identify, measure and visualize the topic analysis and conceptual framing promoted by various politically engaged speakers. For instance, using the coordinating construction it is possible to extract the most important concepts in the conceptual matrix of the lexeme mir "peace" and find the 5 (or more) speakers that most frequently used the concept, as presented in the illustration 1.
The application of the multidimensional graph approach can benefit empirical interdisciplinary type of research in the social sciences and humanities.

Keywords: parliamentary debates, NLP processing, corpus analysis, data integration, social, communication modelling