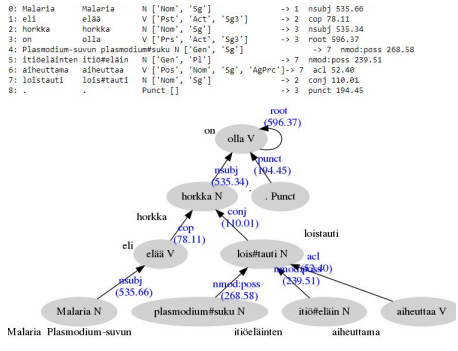
UNIVERSITY OF HELSINKI
DIGITAL HUMANITIES

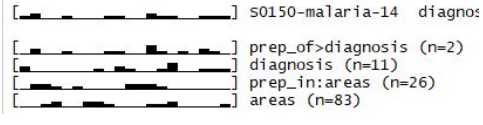
Evaluation of machine-learnt semantic clusters: Emergent concept structures vs. expert taxonomies

Seppo Nyrkkö

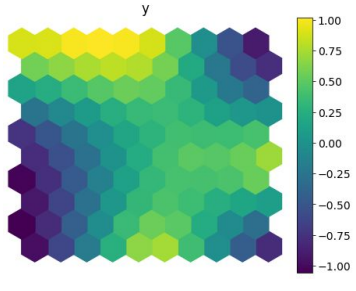
In this work, I study the correspondence between linguistic distributional syntactic analysis of term definitions and hand-made semantic term taxonomies. Automatized syntactic dependency analysis is used for Finnish and English term definitions. Hand-made taxonomies written by human experts are compared to the machine learnt concept structures.



dependency syntax analysis



high-dimensional vector space



emergent concept structures in output space

From Wikipedia texts to syntactic data, and further into high-dimensional vector space

For each term analyzed, the syntactic context is projected into a high-dimensional input vector space. Unsupervised machine learning is applied to the numerical syntactic evidence.

Data analysis methods

SOM - Self-Organizing Map - 2D / 3D lattice representation of the high-dimension input space, model based on brain study
GTM - Generative Topographic Map, a statistical analogy to SOM
PCA, ICA - separation of *feature* and *noise* dimensions
t-SNE - dimension reduction retaining local structures

Emergent conceptual structures can be seen in the grouping of frequent occurrences of highly similar features in an observed dataset. Emergent taxonomies are extracted from the unsupervised clustering. Human-designed semantic web ontologies and taxonomy structures are compared against the emergent clusters. The model can be evaluated even with short dictionary definitions.

Questions answered

Semantic Reasoning:

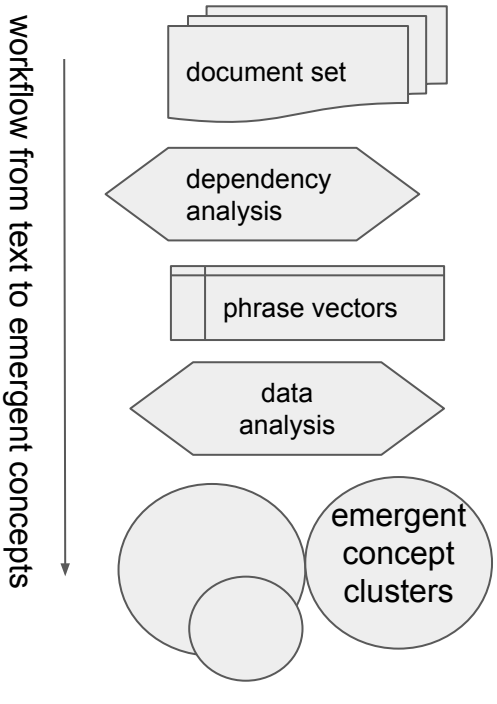
Which semantic features are seen in syntactic distributions?

Hyperonymy	|	Hyponymy
Partonymy	|	Meronymy
Synonymy	|	Antonymy
Causes	|	Effects

Conceptual space model:

Can geometric approach be used in concept representation? Which theoretical feature dimensions are seen in the syntactic evidence?

Colors, Sizes, Physical dimensions
Shapes, Tastes, Symbols



HELSLANG Summer Conference 05 / 2019