Extracting Clusters of Specialist Terms from Unstructured Text

Aaron Gerow
gerow@uchicago.edu
Computation Institute
University of Chicago
Chicago, IL USA

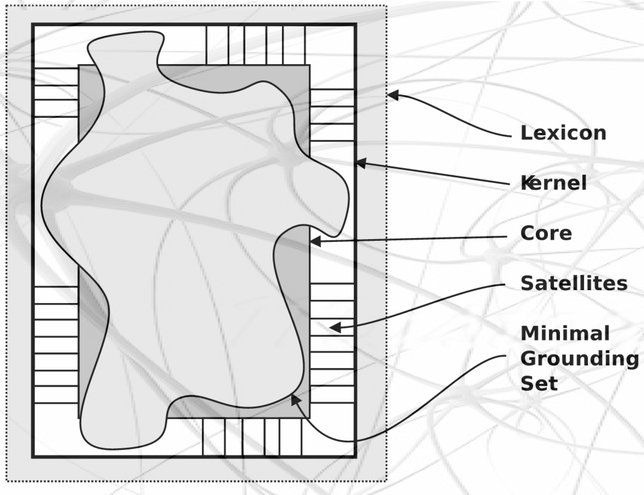
EMNLP 2014: Doha, Qatar

THE UNIVERSITY OF CHICAGO KNOWLEDGE LAB

Abstract

Automatically identifying related specialist terms is a difficult and important task required to understand the lexical structure of language. This paper develops a corpus-based method of extracting coherent clusters of satellite terminology – terms on the edge of the lexicon – using co-occurrence networks of un-structured text. Term clusters are identified by extracting communities in the co-occurrence graph, after which the largest is discarded and the remaining words are ranked by centrality within a community. The method is tractable on large corpora, requires no document structure and minimal normalization. The results suggest that the model is able to extract coherent groups of satellite terms in corpora with varying size, content and structure. The findings also confirm that language consists of a densely connected core (observed in dictionaries) and systematic, semantically coherent groups of terms at the edges of the lexicon.

Peripheral Lexis in Dictionaries



Satellites are likely to be more systematically structured and more coherent than words in the core

Motivation

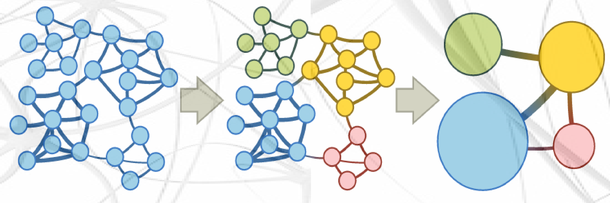
Can terms on the "fringe" of the lexicon be identified and organized from completely unstructured text in a single pass?

Windowed co-occurrence networks naturally push some words to satellite regions

Co-occurrence Network

1. **Construct an edge-weighted network of sentential word-word co-occurrences**
2. **Partition the graph into communities**
3. **Discard the largest community (~90% of words)**
4. **Summarize the remaining clusters by their most internally central words (hub-score)**

Community Detection



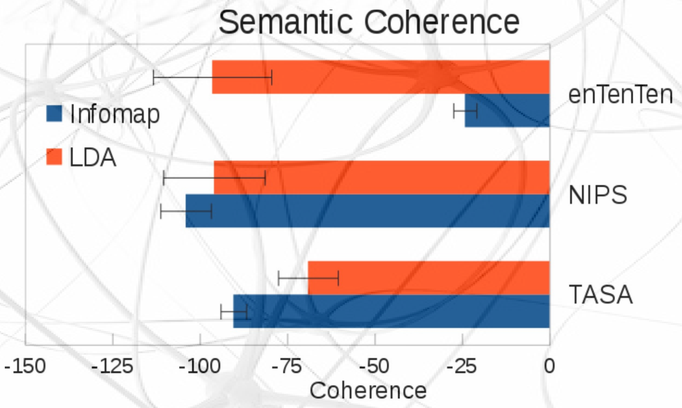
***InfoMap*: By minimizing a description of flow through edges in a graph, densely connected components can be extracted in weighted (and directed) networks.**

Semantic Coherence

$$\frac{1}{n}\sum_{\substack{(w_i,w_j)\in t \\ i<j}}^{n} \log \frac{D(w_i,w_j)+1}{D(w_j)}$$

(Designed to qualify topics by their most likely words)

Results



1. Number of clusters, within-cluster ranking and coherence are all defined intrinsically by the data.

2. Coherent clusters of terms are available without pre-processing or the use of document structure.

3. General language (enTenTen) exhibits more coherent satellite terminology.

4. Extracted terms are signi-ficantly less frequent that "core" words.

TASA		NIPS		enTenTen	
thou	1.00	model	1.00	cortex	1.00
shalt	0.72	learning	0.99	prefrontal	0.88
hast	0.49	data	0.96	anterior	0.41
thyself	0.26	neural	0.94	cingulate	0.33
dost	0.24	using	0.85	medulla	0.28
wilt	0.24	network	0.85	parietal	0.13
canst	0.12	training	0.73	insula	0.13
knowest	0.10	algorithm	0.66	cruciate	0.11
mayest	0.10	function	0.63	striatum	0.11
craven	0.01	networks	0.62	ventral	0.10
peru	1.00	university	1.00	pradesh	1.00
ecuador	0.84	science	0.85	andhra	0.67
bolivia	0.80	computer	0.83	madhya	0.56
argentina	0.67	department	0.74	uttar	0.50
paraguay	0.54	engineering	0.30	bihar	0.21
chile	0.52	report	0.30	rajasthan	0.19
venezuela	0.48	technical	0.29	maharashtra	0.16
uruguay	0.28	institute	0.26	haryana	0.12
lima	0.17	abstract	0.25	himachal	0.10
parana	0.11	california	0.23	arunachal	0.04
clams	1.00	nuclear	1.00	cilia	1.00
crabs	0.87	weapons	0.66	peristomal	0.73
oysters	0.87	race	0.57	stalk	0.62
crab	0.67	countries	0.40	trochal	0.51
lobsters	0.66	rights	0.37	vorticella	0.35
shrimp	0.62	india	0.27	campanella	0.32
hermit	0.50	russia	0.26	hairlike	0.17
mussels	0.27	philippines	0.26	swimmers	0.15
lice	0.23	brazil	0.25	epistylis	0.12
scallops	0.20	waste	0.22	telotroch	0.11

This work was supported by a grant from the Templeton Foundation to the *Metaknowledge* Research Network and grant #1158803 from the National Science Foundation