

M.Tech. Second Stage Poster Presentation

Towards Evaluating Lexico-Semantic Networks

J. Ramanand

Guide - Prof. Pushpak Bhattacharyya

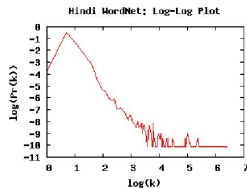
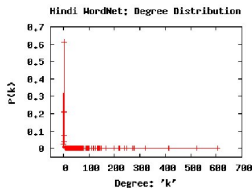
March 12, 2007

Introduction to the Problem

- ▶ Lexico-semantic networks: collection of words, concepts, relations between them
- ▶ Examples: Wordnets, VerbNet, Framenet
- ▶ Use: Natural language applications such as IR, WSD, text summarisation, *etc.*
- ▶ Current status: Increasingly used; Creation of language-specific wordnets (such as Hindi)
- ▶ Key Questions:
 - ▶ What is their quality like?
 - ▶ How to compare two lexico-semantic networks?
 - ▶ How to evaluate and rate?
- ▶ Literature survey revealed very little work has been done in this area
- ▶ Wordnets studied: Princeton English WN, Hindi WN, Marathi WN
- ▶ First step: measure some structural properties about wordnet organisation (degree distribution, avg. shortest path, *etc.*)

Degree Distribution

- ▶ Degree (k) of a wordnet synset: number of semantic relations connecting that synset to other synsets
- ▶ Degree Distribution ($P(k)$): for each k , the proportion of total number of nodes that have degree k
- ▶ A power-law shape was observed for all wordnets (e.g. below)
- ▶ Majority of the nodes have very low degree; a few are 'degree-rich' (e.g.: *city*, *person*, *change*, *vyaktii*)
- ▶ Exponents γ of the power-law give clues about maturity of wordnets



Degree Distribution - Results

Wordnet	Exponent(γ)
English WN (Nouns)	-2.063
English WN (Verbs)	-2.224
Hindi WN	-2.592
Marathi WN	-2.841

Table: Exponents for the Degree Distributions

Some examples:

- ▶ EWN (Nouns): (city,metropolis,urban_center): 664, (law,jurisprudence): 611, (person,individual,someone,somebody,mortal,soul): 400
- ▶ EWN (Verbs): (change,alter,modify): 397, (change): 188, (be): 132
- ▶ HWN: (vyaktii, maanas, shaks, shakhs, ba.ndaa (person)) 607, (karm, karanii, kaam, kaarya, krtya, kaarvaaii, kaarvaahii (action)): 524, (avasthaa, dashaa, haalaat, sthithii, vrttii, suurat, haal, gatii (state)): 414
- ▶ MWN: (vyaktii, maaNus, isama, manushya, paTTThaa, paThyyaa (person)): 626, (karm, krtii, kriyaa, kaam, kaarya, krtya (action)): 546, (avasthaa, sthithii, dashaa, gat (state)): 428

Cluster Coefficient

- ▶ Measures what fraction of neighbours of a node are related to each other
- ▶ *Cluster Coefficient* C_i for a node i (with degree k_i) of a directed graph:

$$C_i = \frac{|E(\Gamma_i)|}{2 \times \binom{k_i}{2}}$$

where Γ_i is the subgraph made of i and its neighbours, $|E(\Gamma_i)|$ is the number of edges of the subgraph, and $2 \times \binom{k_i}{2}$ is the total number of possible edges in Γ_i .

- ▶ Synsets with low degree have high C_i ; High degree synsets have low C_i
- ▶ CC is near 0.5, but seems to be because of clustering in nodes with low degree

Cluster Coefficient - Examples

Wordnet	Avg. Cluster Coefficient
English WN (Nouns)	0.526
English WN (Verbs)	0.632
Hindi WN	0.268
Marathi WN	0.358

Word	Degree	CC
Hamas, Islamic_Resistance_Movement	3	0.667
air_unit	10	0.044
thing	22	0.082
cell	36	0.007
New_Testament	51	0.012
England	85	0.004
baseball	98	0.002
animal_order	102	0.010
military, armed_forces, armed_services, ...	224	0.002
law, jurisprudence	626	0.000

Average Shortest Path Length

- ▶ Low average shortest path even in graphs of these sizes
- ▶ Gives an idea about the diameter of the graphs

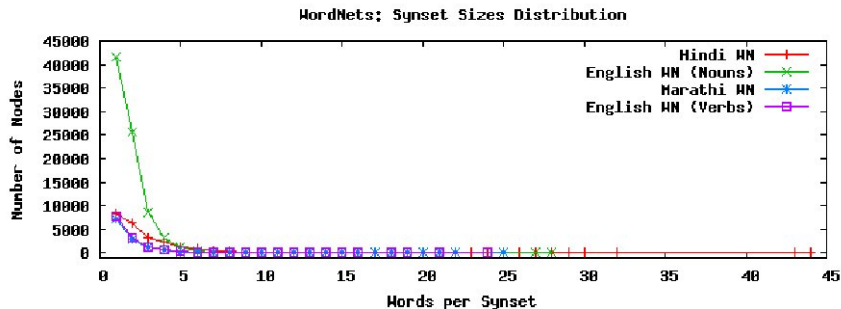
Wordnet	Average Shortest Path	Maximum Shortest Path
EWN (Nouns)	8.878	20
EWN (Verbs)	9.611	27
HWN	4.378	15
MWN	4.255	20

Small World nature of wordnets

- ▶ Low Average Shortest Path Length and High Clustering Coefficient
- ▶ Similar to other natural graphs: web, citation, social networks, ...
- ▶ Seen in all three wordnets - seems to be a language organisation property
- ▶ Hubs and Important nodes
- ▶ Practical consequences: learn from research in other domains

Synset Size Distribution

- ▶ Size of synset: number of words in it
- ▶ Synset sizes distribution: distribution of synonymy;
- ▶ Similar shapes for all wordnets

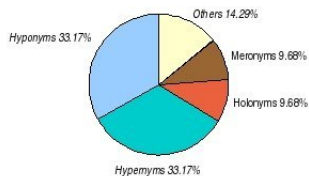


Link Types Distribution

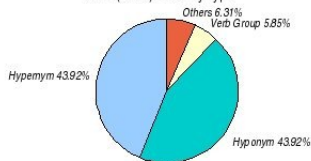
- ▶ Taxonomic relations dominate significantly

10
20
30
40
50
60
70
80
90

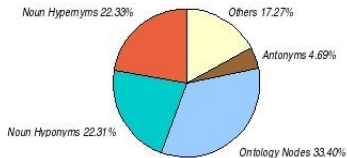
EWN (Nouns) - Links by Type



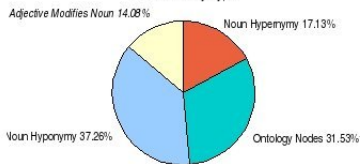
EWN (Verbs) - Links by Type



HWN - Links by Type



MWN - Links by Type



Conclusions

- ▶ Wordnets show 'Small World properties' like other complex networks
- ▶ Similar properties across wordnets; use variations to study quality aspects
- ▶ Only a few degree-rich nodes - a notion of "importance" emerges
- ▶ Exponents in degree distribution may suggest maturity
- ▶ Low Average shortest path; high clustering among low degree nodes
- ▶ Dominance of hypernym/hyponym relations

Future Work

- ▶ Start evaluating wordnet synsets (word collection, gloss, taxonomic links)
- ▶ Evaluation scheme that penalises errors in 'important' nodes
- ▶ Corpora-based evaluation (easier for English than Hindi)
- ▶ Consider sub-graphs of parts-of-speech separately for comparison
- ▶ Bring in non-wordnet networks

References

- ▶ George Miller, Richard Beckwith, Christiane Felbaum, Derek Gross, Katherine J. Miller. *Introduction to Wordnet: an on-line lexical database*. pg 235-244 fo the International Journal of Lexicography 3(4), 1990.
- ▶ Narayan D., Chakrabarty D., Pande P., Bhattacharyya P. *An Experience in building the Indo-Wordnet - A Wordnet for Hindi*. International Conference on Global Wordnet (GWC '02), Mysore, India, 2002.
- ▶ J. Ramanand. *Towards Evaluating Lexico-Semantic Networks*. First Stage Project Report, 2006.
- ▶ Xiao Fan Wang, Guanrong Chen. *Complex Networks: Small-World, Scale-Free and Beyond*. IEEE Circuits and Systems Magazine, First Quarter, 2003.
- ▶ Duncan Watts. *The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2006.
- ▶ Cancho and Sole. *The small world of human language*. Proceedings of the Royal Society, London. Jul 2001.
- ▶ Mariano Sigman, Guillermo A. Cecchi. *Global organization of the Wordnet lexicon*. Proceedings of the National Academy of Sciences of the USA. Vol. 99. Feb 2002.