Unsupervised POS Induction for Bengali

Joydeep Nath, Monojit Choudhury, Animesh Mukherjee,
Chris Biemann and Niloy Ganguly
Contact: monojitc@microsoft.com

objective

Comparison of unsupervised tagset induction techniques for Bengali

Bengali tagset design

Analysis of the word networks to understand the syntactic structure of Bengali

1 acquire a raw text corpus

বাংলা সাহিত্যের মধ্যযুগে বিশেষ এক শ্রেণীর ধর্মবিষয়ক আখ্যান কাব্য মঙ্গলকাব্য নামে পরিচিত। বলা হয়ে থাকে, যে কাব্যে দেবতার আরাধনা, মাহাত্য-কীর্তন করা হয়, যে কাব্য শ্রবণেও মঙ্গল হয় এবং বিপরীতে হয় অমঙ্গল; যে কাব্য মঙ্গলাধার, এমন কি, যে কাব্য যার ঘরে রাখলেও মঙ্গল হয় তাকে বলা হয় মঙ্গলকাব্য। মঙ্গলকাব্য বিশেষ হিন্দু দেবতা যারা "নিম্নকোটি" নামে পরিচিত ছিল তাদের মাহাত্ম বর্ণণায় ব্যবহৃত হত বলে ইতিহাসবিদেরা মনে করেন কেননা এগুলো শাস্ত্রীয় হিন্দু সাহিত্য যেমন বেদ ও পুরাণে অনুল্লেখ্য ছিল।

Target word

Feature word

Feature word, but not function word

Function word, but not feature word

2 extract feature words

The most frequent *m* words are defined as *feature words.*

3 generate context vectors

কাব্য	যে	হয়	PU (।,;)	এই	ও	বলা	...	যার
-2	0	0	3	0	0	0	...	0
-1	3	0	0	0	0	0	...	0
1	0	0	0	0	0	0	...	1
2	0	0	1	0	0	0	...	0

4 construct word network

Words are nodes. The weight of the edge between nodes (words) *u* and *v* is:

$sim_b(u,v) = (1 - \cos(\vec{u}, \vec{v}))^{-1}$
$sim_c(u,v) = \cos(\vec{u}, \vec{v})$

5 cluster the network

Chinese Whispers Algorithm: non-parameterized, random walk based

Agglomerative Hierarchical Clustering: Number of clusters can be decided a priori

6 compute tag-entropy

Measures the goodness of a cluster against a gold standard tagset.

$TE(c) = -\sum[p_i \log p_i + (1-p_i)\log(1-p_i)]$
p_i = fraction of words for which tagi is 1
$MTE = 1/r \sum TE(c_i)$
$WMTE = 1/N \sum |c_i| TE(c_i)$
r = number of clusters
N = number of nodes in the network

I. Topological properties of word networks

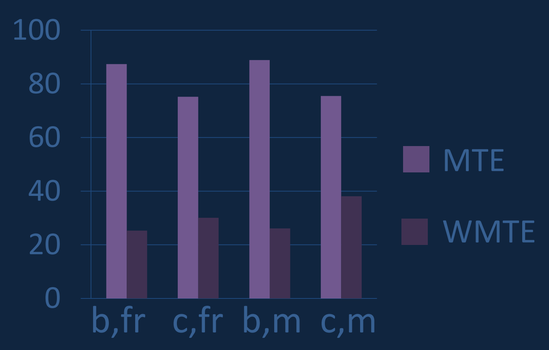
Property	Nature	Conclusion
Degree distribution	Power-law with exponent -1	Hierarchical organization of ambiguity classes
Clustering coefficient	0.53 (high positive correlation with degree)	Frequent words are ambiguous; existence of large clusters
Cluster size	Power-law with exponent -1.02	The fractal nature of the networks

II. Tag-entropy based analysis

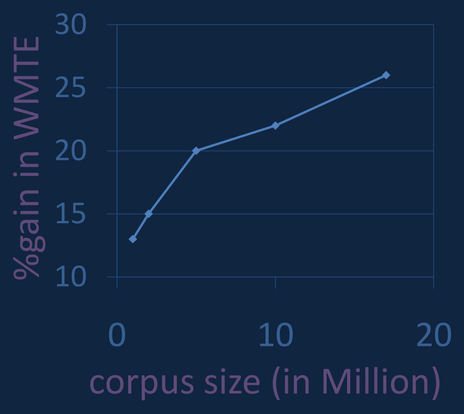
Baseline: All the words in same cluster

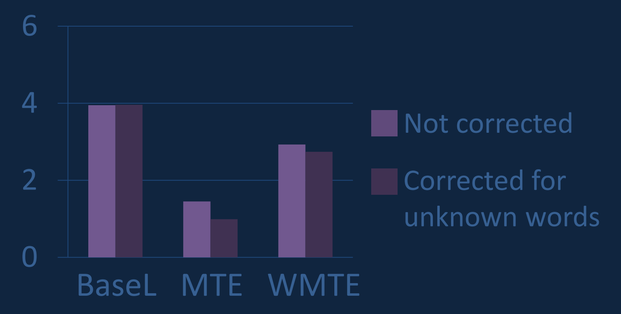
Metric: % gain in MTE or WMTE

$$\frac{\text{Baseline entropy} - \text{(W)MTE}}{\text{Baseline entropy}}$$



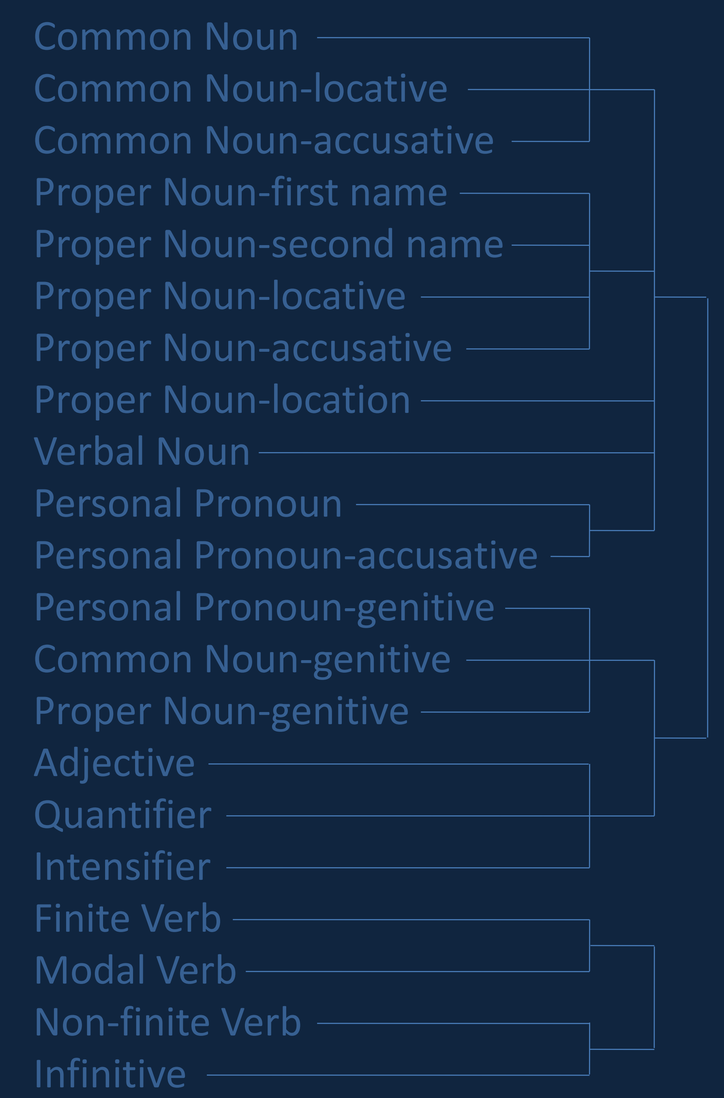
%gain in MTE & WMTE for various graph construction & agglomerative hierarchical clust.





Tag entropies for n = 17M, m = 50, using sim_b & CW algorithm

III. Linguistic Analysis



We observe no distinctions between the distributions of singular and plural nouns.

Example clusters are available at
http://banglaposclusters.googlepages.com/home

Example Clusters:

Cluster 1:
Proper Nouns
buddhabAbu
saurabha
rAkesha

Cluster 2:
Noun-genitive
golamAlera (of problem)
dAbira (of right)
phalera (of result)

Cluster 3:
Quantifiers
sAtaTi (seven)
anekaguli (many)
3Ti (three)

Cluster 4:
Noun-locative
adhibeshane (during the session)
dalei (in party)
baktritAYe (in speech)
bhAShaNe (in speech)

Cluster 5:
Infinitives
bhAbte (to think)
khete (to eat)
jitate (to win)