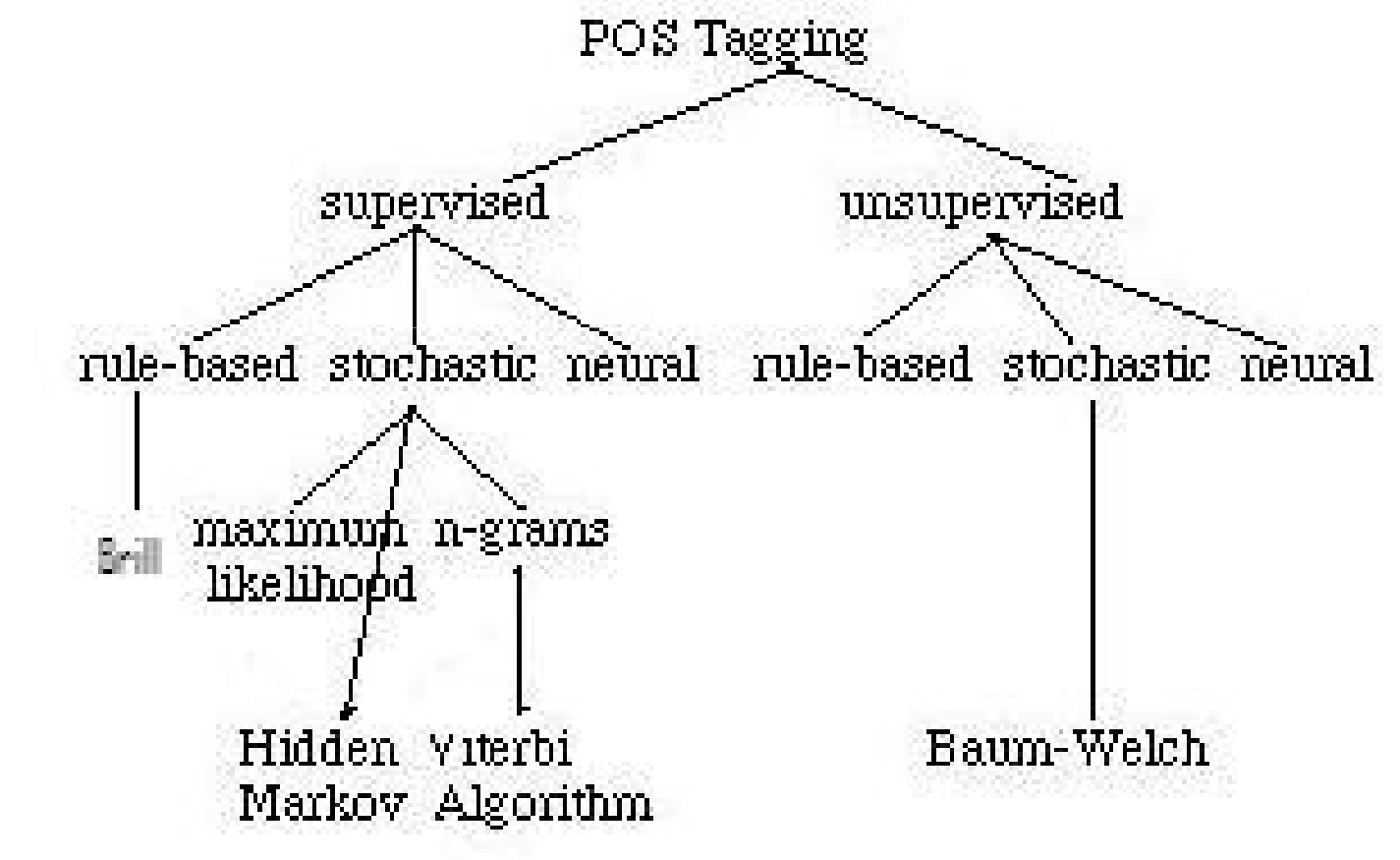
An online semi automated POS tagger for Assamese

Presented By:
Pallav Kumar Dutta
Indian Institute of Technology Guwahati

Topics

- Overview, Existing Taggers & Methods
- Basic requirements of POS Tagger
- Our Approach
- Future work
- References

Overview of POS tagging



Existing Taggers

- Rule Based - uses a set of rules & contextual info to tag a word.
- Stochastic Taggers – Uses frequency, probability to tag a word.

Statistical Approaches

- Hidden Markov Model (HMM)
- Maximum Entropy Model (ME)
- Conditional Random Field (CRF)
- Memory Based Learning (MBL)

Basic Requirements of POS tagger

- Tag Set
- Corpus

Indian Langage tagset

- Very few are publicly available and motivated by specific research agenda
- IL tagset – developed at IIIT Hyderabad
- This tagset has adopted a very coarse structure in linguistic analysis, leading to a very flat structure capturing
  - Avoiding singular vs plural distinction, tense distinctions, and almost all the case markers except the Locative case
  - Tag “PREP” used for POSTP also.

Assamese Tagset

CC	Conjunct
NN	Noun
NVB	Noun in Kriyamul
PRP	Pronoun
PUNC	Punctuation
QF	Quantifier
QFNUM	Number Quantifier
QW	Question Word
RB	Adverb
RP	Particle
SYM	Symbol
UH	Interjection
INTF	Intensifier
JJ	Adjective
NEG	Negative
VAUX	Verb Auxiliary
VAUXN	Verb Auxiliary Negative

Assamese Tagset ..

VFM	Verb Finite Main
VFMN	Verb Finite Main Negative
VNF	Verb Nonfinite
VNFN	Verb Nonfinite Negative
JVB	Adjective in Kriyamul
POSTP	Post Position
DJV	Deadjectival Verb
DNV	Denominal Verb
DNVN	Denominal Verb Negative
DVJ	Deverbal Adjectival
DVJN	Deverbal Adjectival Negative
DVN	Deverbal Nominal
DVNN	Deverbal Nominal Negative
DVR	Deverbal Adverbial
DVRN	Deverbal Adverbial Negative
FW	Foreign Word
OVB	Onomatopoetic Word in Kriyamul
OW	Onomatopoetic Word

TAGSET	TYPE	FORMS	FEATURES	COMMON FEATURES
NN	NC NP NLOC NM NA NCL		NUM, PM, CL, FEM, CASE, COMP, POSTP, EMP	CLS, FEM, CASE, DEG, POSTP, EMP
PRP	PERS	1SG 1SG.NOM 1PL 1PL.NOM 2_1 2_2 2_3 2_1SG 2_2SG 2_3SG 2_1SG.NOM 2_2SG.NOM 2_3SG.NOM 3_1 3_2 3_3F 3_3M 3_1.SG 3_2.SG 3_3F.SG 3_3M.SG 3_1.SG.NOM 3_2.SG.NOM 3_3F.SG.NOM 3_3M.SG.NOM	CLS, CASE, DEG, POSTP, EMP	
	REL RFLX REC DEM IND			
FW	FW		CL, CASE, DEG, POSTP, EMP	
JJ	JJ		EMP	EMP
RB	RB JJ		RLZ, EMP	
VFM	V		CAUS, ASP, TNS, PRS, MOOD	CAUS, ASP, TNS, PRS, MOOD

VFMN	V	NEG, CAUS, ASP, TNS, PRS, MOOD	
VNF	V	NF SUFFIXES (primary, secondary, peripheral)	NF SUFFIXES (primary, secondary, peripheral)
VNFN	V	NEG, NF SUFFIXES (primary, secondary, peripheral)	
VAUX	MOD ASPL PASS	ASP, TNS, PRS, MOOD	ASP, TNS, PRS, MOOD
VAUXN		NEG, ASP, TNS, PRS, MOOD	
DVJ	V	JLZ	JLZ
DVJN	V	NEG, JLZ	
DVR	V	JLZ, RLZ	JLZ, RLZ
DVRN	V	NEG, JLZ, RLZ	
DVN	V	NLZ, CL, CASE, POSTP	NLZ, CL, CASE, POSTP
DVNN	V	NEG, NLZ, CL, CASE, POSTP	
DNV	NN	VLZ, ASP, TNS, PRS, MOOD	VLZ, ASP, TNS, PRS, MOOD
DNVN	NN	NEG, VLZ, ASP, TNS, PRS, MOOD	
DJV	JJ	VLZ, ASP, TNS, PRS, MOOD	
DOV	OW	VLZ, ASP, TNS, PRS, MOOD	
NVB	NN		
JVB	JJ		
OVB	OW		
CC	COMP		
RP	RP		
QF	QF	NN, CASE, EMP	
QFNUM	NUM	CL, EMP	
INTF			
POSTP	POSTP	EMP	
QW	QW	CL, CASE, DEG, POSTP, EMPH	
UH			
PUNC			
SYM			
NEG			
OW			

Corpus Generation

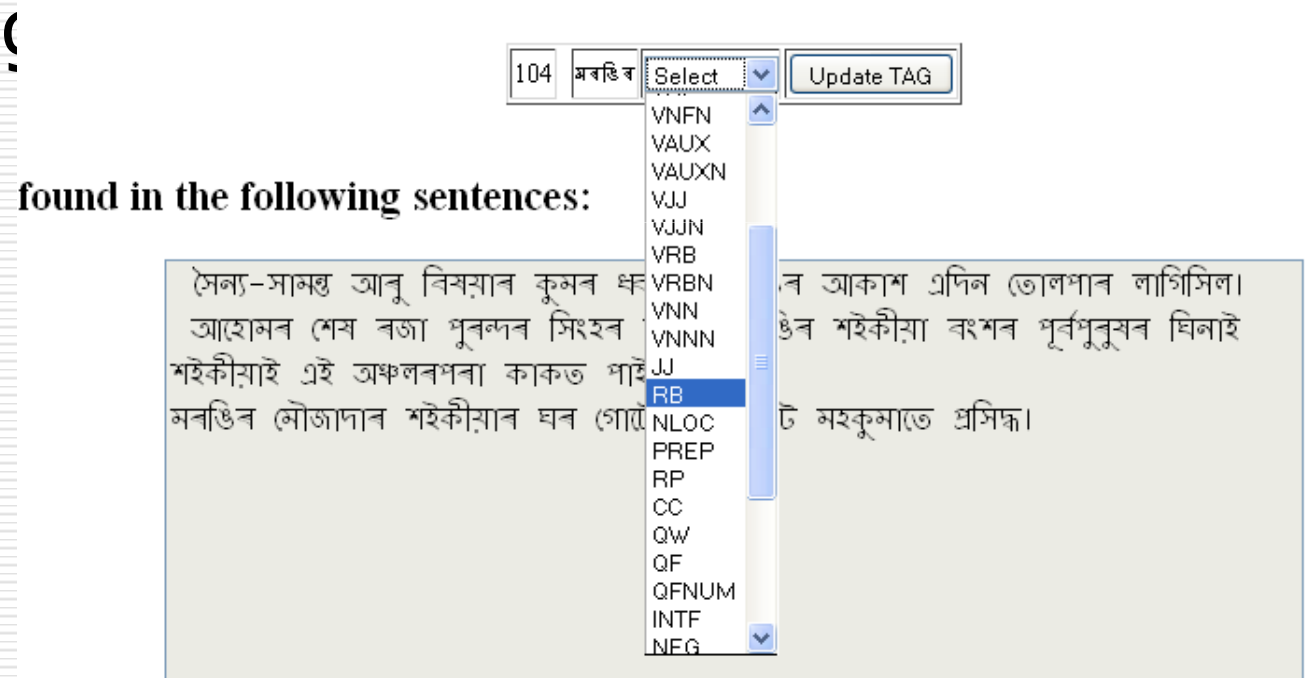
- Unicode supported corpus
- Training corpus of 1860 sentences (20414 tokens/words)
- Collected from short stories from famous novels, news paper etc.

Our Approach of POS Tagger

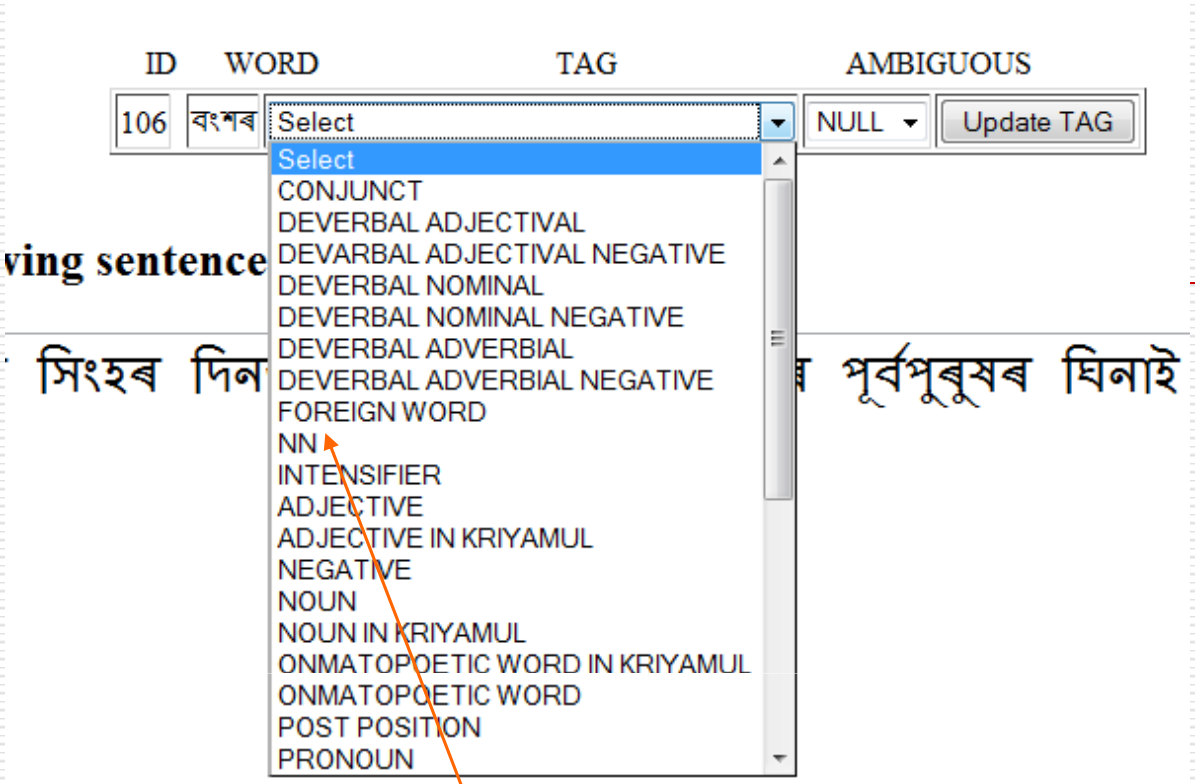
- All statistical approaches require a huge training set.
- Rely on pattern matching of new instances with what is already available with native speakers providing verification.
- We have observed that providing simple aids to human improves the productivity tremendously.

POS tagger

- The online tagg



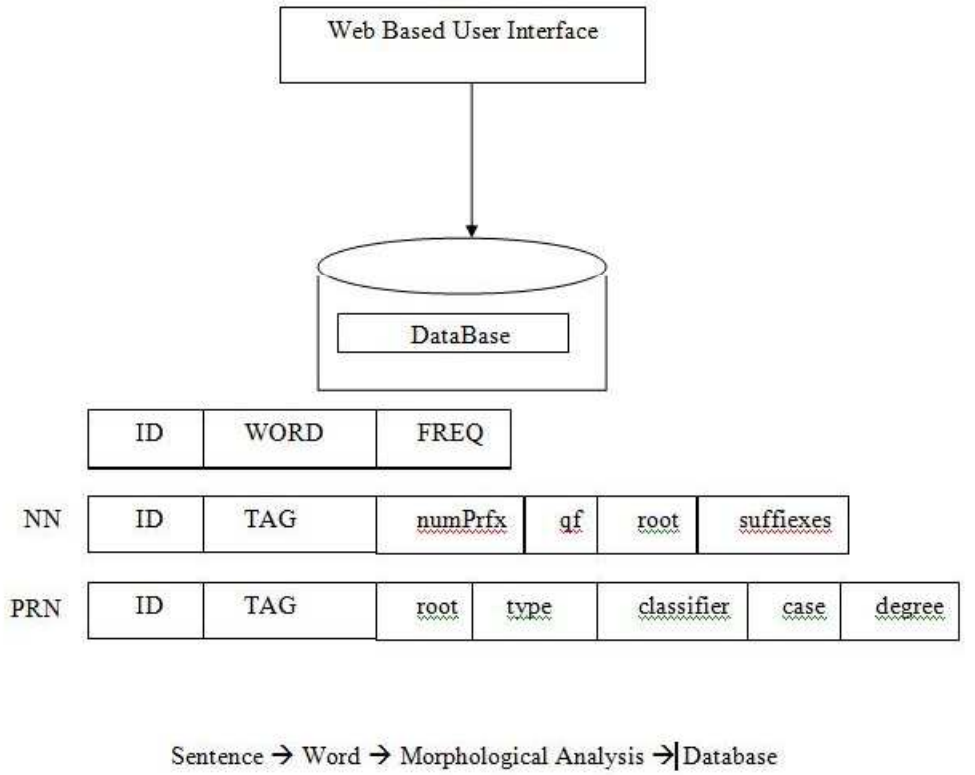
A simple tagger



After selecting "NN", the following detail analysis page appears

POS Tagger

- Uses Database.



Observations

- Intra & inter domain effects.
- Sparse data problem
- Tested with NLTK

Future Work

- We will seek to generalize this process so that it becomes applicable to other NE Languages and we aim to create such a system for at least one more language like Bodo.

- Bodo training corpus contains 1571 sentences with 19658 tokens.

References:

- Fahim Hasan, N UzZaman, Mumit Khan, 2009. "Comparision of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages", [22] N Saharia, D das, U Sharma, Jugal Kalita, " Part of Speech Tagger for Assamese Text", Proceedings of the ACL-IJCNLP 2009 Conference, pp. 33-36.
- Baskaran, S Et al., 2008. "Designing a common POS-Tagset Framework for Indian Languages", Proceedings of the 6th Workshop on Asian Language Resources (ALR 6), Hyderabad, pp. 89
- Chirag Patel and Karthik Gali, 2008. "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, Hyderabad, India, pages 117–122.
- A. Dalal, K. Nagaraj, U. Swant, S. Shelke and P. Bhattacharyya, 2007. "Building Feature Rich POS Tagger for Morphologically Rich Languages: Experience in Hindi ". ICON.
- Sandipan Dandapat, Sudeshna Sarkar, Anupam Basu, 2007. "Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario" Proceedings of the ACL 2007 Demo and Poster Sessions, pages 221–224.
- Himanshu Agrawal, 2007. "POS Tagging and Chunking for Indian Languages" Proceedings of IJCAI Workshop on Shallow Parsing for South Asian Languages, Hyderabad.
- Avinesh PVS and Karthik G, 2007. "Part-Of-Speech Tagging and Chunking Using Conditional Random Fields and Transformation Based Learning", Proceedings of IJCAI Workshop on Shallow Parsing for South Asian Languages, Hyderabad.

References (contd.)

- Delip Rao and David Yarowsky, 2007. "Part Of Speech tagging and Shallow Parsing for Indian Languages", Proceedings of IJCAI Workshop on Shallow Parsing for South Asian Languages, Hyderabad.
- T. N Vikram and Shalini R Urs, 2007. "Development of prototype Morphological Analyzer for the South Indian Language of Kannada", Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers. PP. 109-116.
- Ms Lilabati Saikia Bora ,Asamiya Bhasar Ruptattva- First Edition, January 2006,published by M/s Banalata, Panbazar Guwahati-1. (ISBN: 81-7339-466-0)
- Satyanath Borah- Bahal Vyakaran, April 2006 Ed, published by B.C. Barua on behalf of Gopal Barua Agency, S.B Road, Guwahati-1
- Shrivastava, Agrawal, Mohapatra, Singh and Battacharya, 2005. "Morphology based Natural Language Processing tool for Indian languages", paper presented in the 4th Annual Inter Research Institute Student Seminar in Computer Science (IRISS05), April, IIT Kanpur. (www.cse.iitk.ac.in/users/iriss05/m_shrivastava.pdf)
- Sirajul Islam Choudhury, L. Sarbajit Singh, S. Borgohain and P. K. Das, , 2004 "Morphological Analyzer for Manipuri: Design and Implementation", Applied Computing, pp. 123- 129.
- Prof. G. C. Goswami, Asamiya Vyakaran Pravesh- Second Edition, April 2003, published by M/s Bina Library, Guwahati-1.
- Eric Brill, 1992. "A simple rule-based part of speech tagger", Proceedings of the Third Conference on Applied Natural Language Processing, pp. 152—155.
- Akshar Bharati, Vineet Chaitanya, Rajeev Sangal "Computational Linguistics in India: An overview"
- http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf
- http://ucrel.lancs.ac.uk/claws/
- http://www.nltk.org

Thanks