

Transformation based Brill's tagger or HMM? Which technique to consider for Bangla POS tagging?

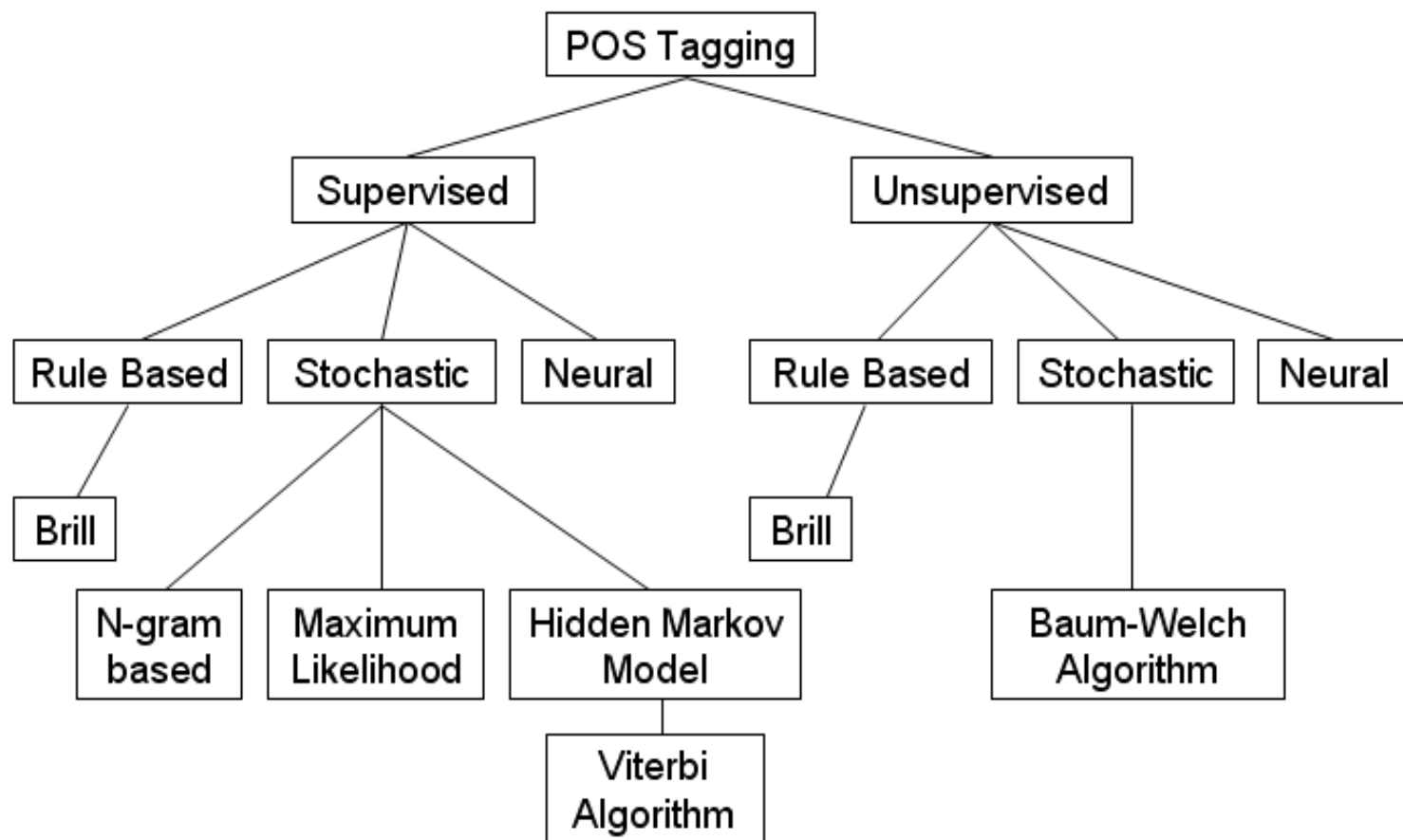
Naushad UzZaman, Fahim Muhammad Hasan & Mumit Khan
Center for Research on Bangla Language Processing (CRBLP)
BRAC University, Bangladesh

Workshop on Morpho-Syntactic by
The School of Asian Applied Natural Language
Processing for Linguistics Diversity and Language
Resource Development (ADD), Bangkok, Thailand
March 6 – 14, 2007

Introduction

- POS tagging - assigning tags to words in text
- Useful for parsing, TTS, IR, semantics etc
- Bangla - 200,000,000+ native speakers
- Lacks significant research in NLP
- Very few works in POS tagging for Bangla
- Even fewer for Bangladeshi Bangla

POS tagging models



Methodology

- Hidden Markov Model (HMM) tagger
 - $P(\text{word} \mid \text{tag}) * P(\text{tag} \mid \text{previous } n \text{ tags})$
- Transformation based Brill's tagger
 - guess the tag of each word, then go back and fix the mistakes.

POS Tagsets

- Bangla Tagset with 3 levels
 - Level 1 (12 tags)
 - Level 2 (41 tags)
 - SSF format (26 tags)

Training Corpus and Test Set

- Training Corpus
 - Prothom-alo newspaper corpus, around 5,000 words tagged with level 1 (12 tags) and level 2 (41 tags)
 - Data provided in the SPSAL contest, around 25,000 words tagged with SSF format (26 tags)
- Test Set
 - 85 sentences, 1000 tokens from prothom-alo corpus
 - 400 sentences, 5225 tokens from corpus of SPSAL contest

Performance of POS taggers for Bangla

Training corpus (no of tokens)	Tagset size (no of tags)	HMM Bangla (percentage)	Brill's Bangla (percentage)
4484	12	45.6	71.3
4484	41	46.9	54.9
25456	26	63.6	69.6

Performance of POS taggers for English

Training corpus (no of tokens)	Tagset size (no of tags)	HMM English (percentage)	Brill's English (percentage)
4042	Brown tagset	70	67.5
30017	Brown tagset	83.1	78.8
400017	Brown tagset	89.7	88.5

Results

- Low Accuracy
 - Small training corpus
 - Not well-balanced corpus
- Transformation based Brill's tagger performs better than HMM for Bangla

Summary and Conclusion

- Experimented Bangla POS taggers
 - HMM and Transformation based Brill's tagger
 - Three different tagset (12, 26 and 41 tags)
 - Different sized corpus (4,000 to 25,000)
- Transformation based Brill's tagger performed better than HMM
- Rule based approach works better for Bangla
- Bangla is derived from a rule-based language, Sanskrit, which can be described by rules.

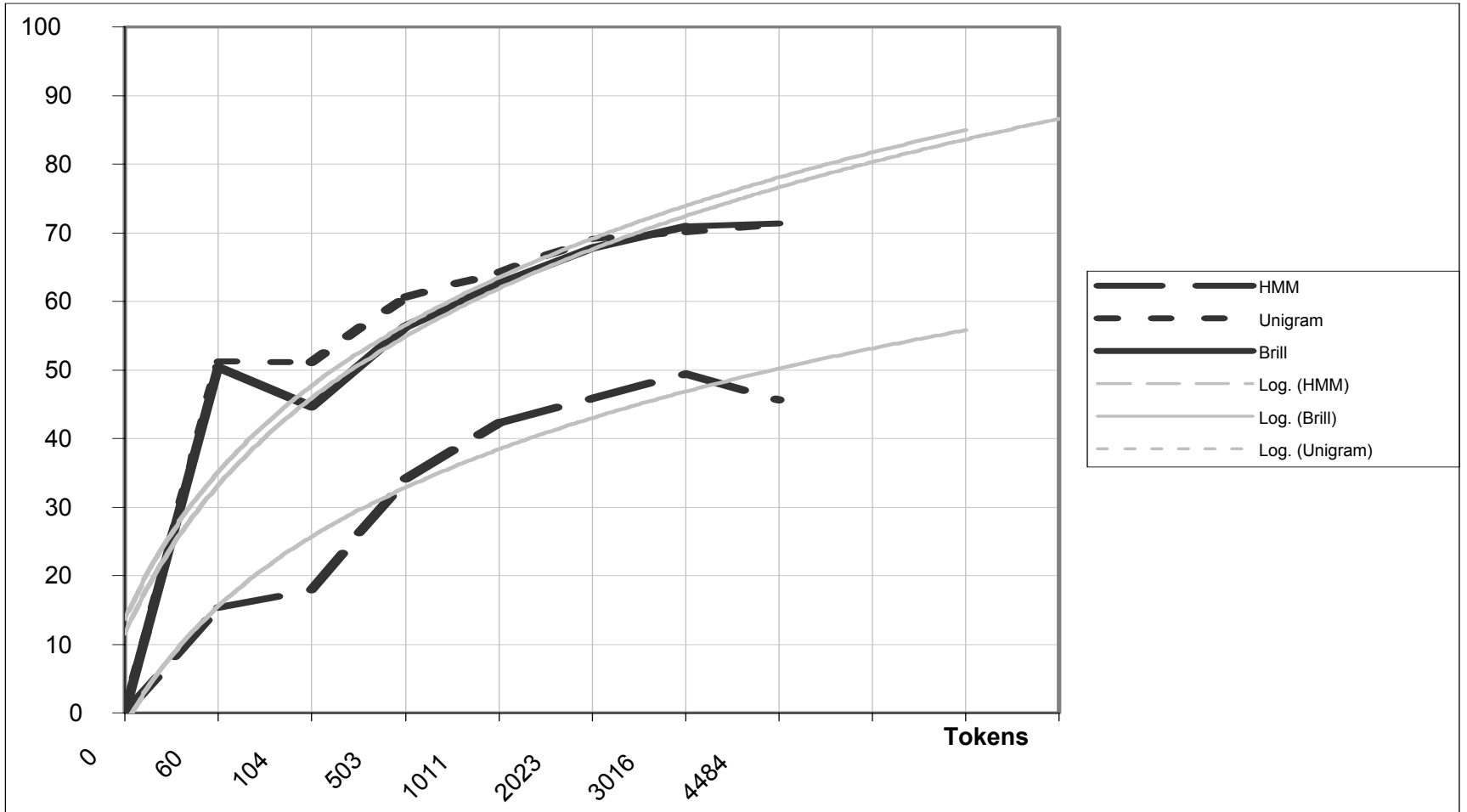
Future Work

- Experiment on larger corpus
- Try unsupervised approaches
- Try Hybrid solution
- Try Transformation based Brill's tagger for other South Asian languages

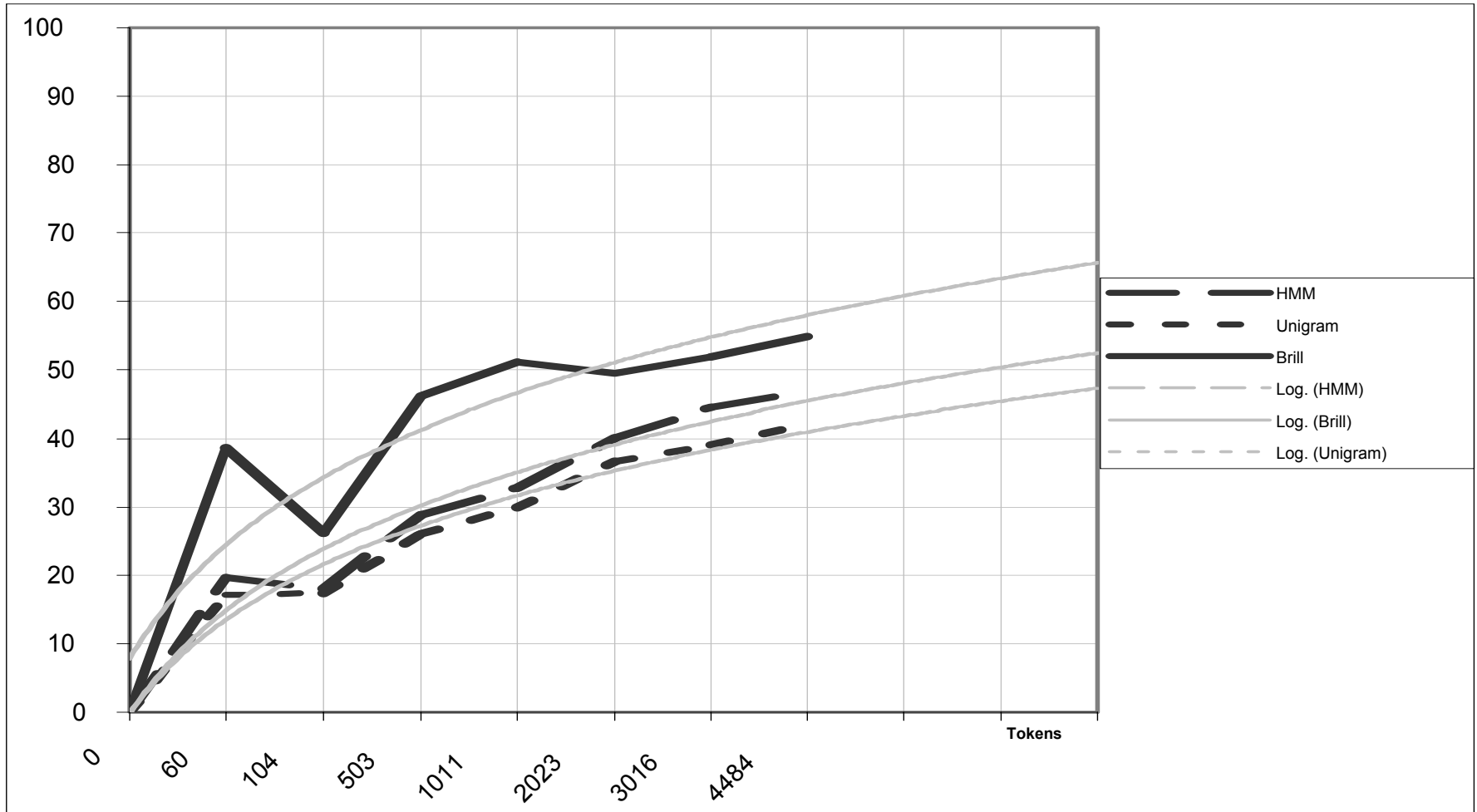
SSF format

- We used the 26-tags tagset provided for the SPSAL contest for our experiments with Bangla. The tagset is based on the Penn-Treebank tagset and consists of the following tags: Noun (NN), Proper Noun (NNP), Pronoun (PRP), Verb Finite Main (VFM), Verb Auxiliary (VAUX), Verb NonFinite Adjectival (VJJ), Verb NonFinite Adverbial (VRB), Verb NonFinite Nominal (VNN), Adjective (JJ), Adverb (RB), Noun Location (NLOC), Postposition (PREP), Particle (RP), Conjunct (CC), Question Words (QW), Quantifier (QF), Number Quantifiers (QFNUM), Intensifiers (INTF), Negative (NEG), Compound Common Nouns (NNC), Compound Proper Nouns (NNPC), Noun in kriya mula (NVB), Adj in kriya mula (JVB), Adv in kriya mula (RBVB), Interjection (UH), Special : Not classified in any other (SYM) (SPSAL, 2007).

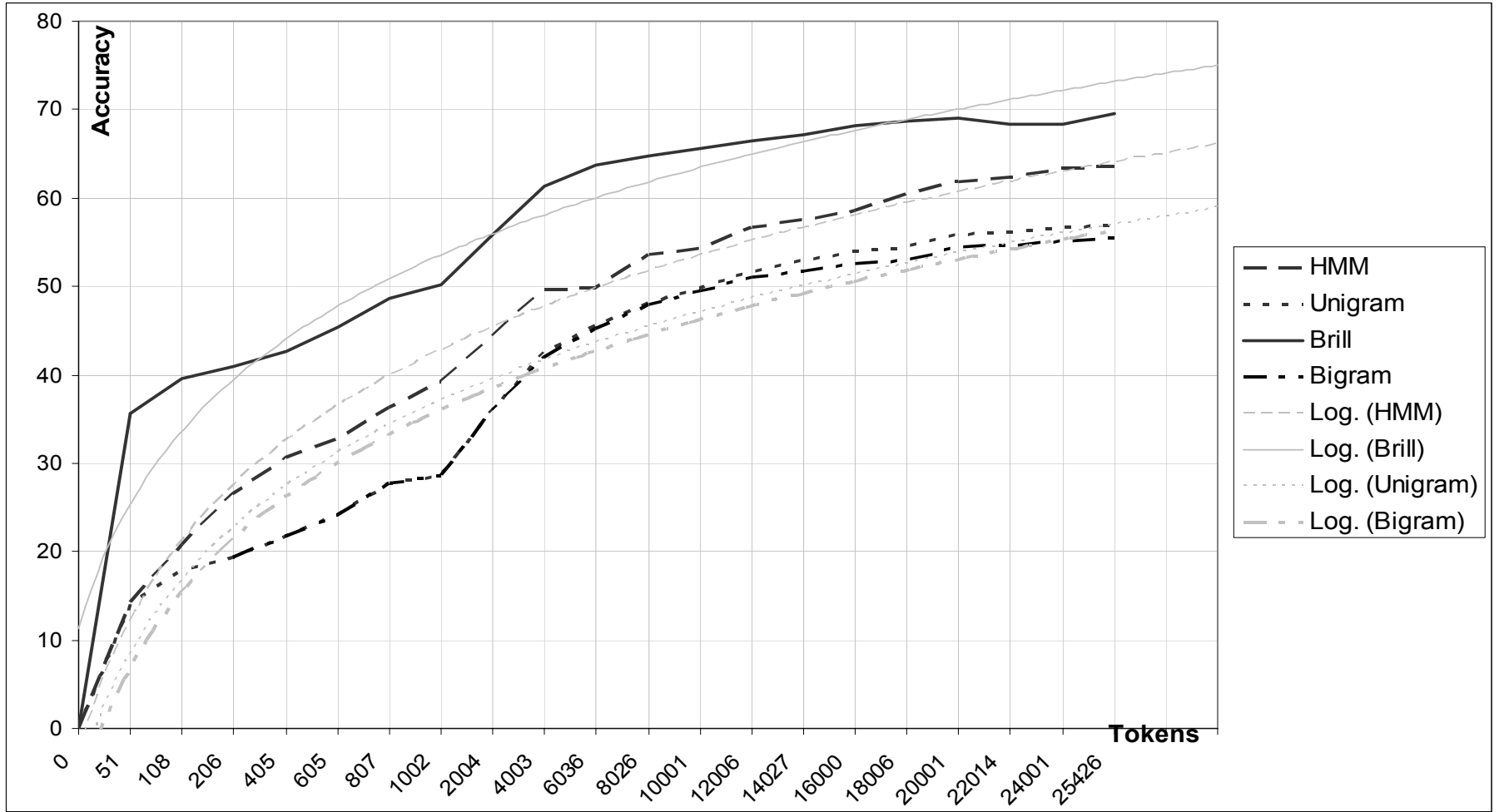
Test data: 85 sentences, 1000 tokens from the (Prothom-Alo) corpus;
Tagset: Level 1 Tagset (12 Tags)]



Test data: 85 sentences, 1000 tokens from the (Prothom-Alo) corpus;
Tagset: Level 2 Tagset (41 Tags)



Test data: 400 sentences, 5225 tokens from the corpus by SPSAL contest;
Tagset: SSF Tagset (26 tags)



Test data: 22 sentences, 1008 tokens from the Brown corpus; Tagset: Brown Tagset]

