



Tagset and Tagging Bangla Corpus

Altaf Mahmud and Mumit Khan

BRAC University

Bangladesh

Introduction

- A Penn Treebank inspired tagset (54 tags)
- A News corpus with approx. 1 million tokens
- Automatic tagging using a small manually tagged training set
- Currently manually tagging the 100k token PanL10n shared corpus derived from the Brown Corpus

A bit of history ...

- **Tagsets are hard!**
- Started with a Hindi tagset as the baseline (very little linguistic focus, mostly ad-hoc)
- Moved to using EAGLE framework (designed for multiple languages, with strong linguistic focus)
- Based current one on Penn Treebank because of our interest in creating a Bangla treebank.
- “Minimalism” a design goal

Tagset design challenges

- Only ad-hoc tagset development possible without thorough linguistic review (see EAGLES guidelines for example)
- Domain independence may simply be too ambitious for (human and linguistic) resource starved languages
- Tagging narrow domain corpus does not allow proper evaluation of the tagset

Tagset overview

	Level 1	Level 2
1	Noun	Common, Proper, Compound Common, Compound Proper, Verb Root, Temporal, Locative
2	Pronoun	Personal Pronoun
3	Vocatives	-
4	Adjective	Simple, Verb Root
5	Verb	Main Finite, Non-finite Nominal, Non-finite Conditional, Nonfinite Perfective, Non-finite
6	Adverb	-
7	Conjunction	Coordinating, Compound Coordinating, Suspicion, Compound Suspicion, Subordinating, Compound Subordinating
8	Numbers	Cardinal Numbers
9	Adposition	--

Tagset overview - continued

	Level 1	Level 2
10	Interjection	-
11	Particle	-
12	Determiner	Simple, Singular
13	Question Word	-
14	Quantifier	-
15	Suffixes	Adpositional, Plural, Accusative, Possessive, Determinative, Adverbial, Particle
16	Foreign Word	-
17	Symbol	-
18	Taka	-
19	Punctuation Marks	...

Tagset details - Noun

Level 2	Tag	Example
Common	NN	মানুষ, পানি
Proper	NNP	মতিউর, অে ঙ্কর, ঢাকা, চট্টগ্রাম, শনিবার
Compound Common Noun	NNC	ছেলে/NNC মেয়ে/NN, স্বরাষ্ট্র/NNC মন্ত্রনালয়/NN
Compound Proper Noun	NNPC	আব্দুর/NNPC রহমান/NNPC বিশ্বাস/NNP
Verb Root	NNV	ে ঙ্কল, পান
Temporal	NNT	গতকাল, আগামীকাল, আজ
Locative	NNL	উপর, নিচ, আগে

Tagset details - Verb

Level 2	Tag	Example
Main Finite Verb	VB	করি, কর, করে, করাই, করলাম, করলে, করেছিস, করব, করাব
Nonfinite Nominal	VBM	করা, করাতে ম্ পরা, পরাতে ম্
Nonfinite Conditional	VBC	করলে, করালে
Nonfinite Perfective	VBT	করে, গিয়ে
Nonfinite	VBF	করতে, করাতে

Tagset details - Suffixes

Level 2	Tag	Example
Adpositional	SFON	এ, য়, তে
Plural	SFL	রা, এরা, গুলি, গণ
Accusative	SFAC	কে, রে, এরে, দিগকে, দিগেরে
Possessive	SF\$	এর, দের
Determinative	SFDT	টা, টি
Adverbial	SFRB	ও, ই
Particle	SFRP	ত

Comparison

- IIIT-Hyd tagset of 26 POS tags designed for Indian languages for SPSAL-2007 contest:
shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf
- IIT-Kgp tagset of 40 tags for Bengali language:
www.mla.iitkgp.ernet.in/Tag.html
- CRBLP tagset with 42 general POS tags, and 9 other tags (punctuation marks and symbols), total 51 tags

Comparison with IIT-Hyd

- No auxiliary verbs in Bangla, so there is no VAUX category in our tagset
- Instead of Verb NonFinite Adjectival and Verb NonFinite Adverbial, we have other categories in NonFinite Verb for Bangla. Have Verb NonFinite Nominal category as well
- All question words subsumed under QW; relative words are tagged as simple type. Separate category for each question word type and subsume the corresponding relative types under that category

Comparison with IIT-Kgp

- No description or linguistic analysis
- Size of their tagset and the primary categorization is closely comparable with our proposed tagset
- A (preliminary) comparison study shows that our tagset precisely distributes all of the required syntactical categories
- It can be expected that by using our tagset, the precision level will be as high as their recent experiments

Strengths and weaknesses of CRBLP tagset

- Strengths
 - Documented and linguistically analyzed
 - Battle-tested on a news corpus
- Weaknesses
 - Large tagset
 - 42 general POS tags
 - 9 other POS tags (punctuation marks and symbol)
 - Total 51 tags
 - “Real” tagging experience is still in its infancy

Corpus tagging methodology

- 4-member team– 2 taggers, tagset designer, independent reviewer (linguist)
- Each tagger tags a separate section, and reviews each other's work
- Tagset designer serves as the tie breaker
- Independent reviewer who periodically looks at completed chunks

Corpus tagging challenges

- Linguistic ambiguity difficult to avoid – steep learning curve beyond the basic few tags
- Bangla's (semantic) case system causes significant problems for the taggers
- Reviewing is a difficult process without having computational linguistic interest (*the battle of the theoretical linguists and taggers rage on ...*)

Tagging evaluation

- Have not chosen a formal framework for evaluation, mostly ad-hoc at this point
- Tagging the parallel corpus would provide guidance on how to evaluate properly

Summary

- Second generation tagset designed for syntactic bracketing (Treebanks)
- Documented, with examples
- Tagged experience with newspaper corpus
- Beginning to tag a balanced corpus – very different experience
- First known linguistically analyzed and documented tagset for Bangla

Future work

- Complete the tagged parallel corpus
- Review the tagged news corpus
- Thorough linguistic review of the tagset after both are done, and iterate once more