TOWARDS POS TAGGING OF MAITHILI

Presented by:

Dr. Arun Kumar Singh

GOALS OF THIS PRESENTATION

- About Maithili Language

- Introducing BIS POS Tagset for Maithili

MAITHILI

- **Maithili, an Indo-Aryan language, is the 16th largest language of India (recognized as a scheduled language in 2003).**
- **It is spoken by a total of about 22 million people in the eastern and northern regions of the Bihar, India and the south eastern plain, i.e. the tarai of Nepal.**

- **Linguistically, Maithili is an inflectional language having relatively free word-order, yet the predominant word order is subject-object-verb (SOV).**

राम आम खाइत अछि

Ram mango eat-hab be-prs

- **Technically, it is one of the resource-poor languages and also relatively less poor than other major Indian languages.**

PART OF SPEECH (POS) TAGGING

- Part of Speech (POS) tagging is the process of assigning correct part of speech to each word in the given context.

- Though POS tagging for major Indian Languages in general has been done in recent years, Maithili has not been explored yet. As per our knowledge no attempt has been made for developing a POS tagger for Maithili.

- The first effort has been made in the project Linguistic Data Consortium for Indian Languages (LDC-IL).

- The BIS POS tag set for Maithili has been recently designed by the LDC-IL.

BUREAU OF INDIAN STANDARDS (BIS) SCHEME

- **Principle 1 - Tag set should be generic with layered approach.**
- **This means that it should be applicable to any language (with flexibility).**
- **The layered approach is important as the task of capturing all the linguistic information can not be done at one level.**

- **Principle 2- Hierarchy within each layer**
- **Principle 3- Hierarchy at the POS level too.**
- **The finer information, i.e. morph features, syntactic and semantic roles etc. can be captured at other layers.**

- **Based on these principles a super tag set for Indian Languages was prepared and then tag sets for each language.**

- **Keeping in view all the above principles, LDC-IL has designed a BIS Tagset for Maithili.**

MAITHILI POS TAG SET

- We will first discuss the categories included for Maithili. Maithili tag set has total 11 categories.

1. Noun (N)
2. Pronoun (PR)
3. Demonstrative (DM)
4. Verb (V)
5. Adjective(JJ)
6. Adverb (RB)
7. Postposition (PSP)
8. Conjunction(CC)
9. Particle (RP)
10. Quantifiers(QT)
11. Residual (RD)

NOUN (N)

- Common noun (NN) - पोथी, कलम, पंडित
- Proper noun (NNP) – अरुण, दिनेश, अतुल
- Nloc(NST) - आगू, पीछू

➢ The annotation of Proper Noun is difficult as in Indian languages, NNP is not easily identifiable unlike English. Hence, to recognize it at later level, we are tagging all the tokens that come under one chunk as NNP, NNP, NNP. For example,
पंडित /NNP अरुण/NNP कुमार/NNP

- **Nloc(NST) – These are spatio-temporal nouns but can function differently. For example,**
- **राम आगू/NST गेल**

  Ram outside go-pst

- **घरक आगू/PSP मंदिर अछि**

  house-gen outside temple be-prs

PRONOUN (PR)

- Personal(PRP)- तौं, अहाँ, हम, ई, एहि, ओकर,तकर
- Reflexive(PRF)- अपना, अपने, स्वयं, खुद
- Relative(PRL)- जे, जिनका, से, सेहो, जेहो, सैह
- Reciprocal(PRC)- आपसमे, परस्पर, एक-दोसरकेँ
- Wh-wordPRQ)- के, कथी, की, ककर
- Indefinite(PRI)- केओ, किछु, कतए, कतो

DEMONSTRATIVE

- Demonstrative has been kept as a distinct grammatical category as they do not behave like pronouns, i.e. they are not variables and they do not take nominal inflections. It has four types:
- Deictic(DMD)- अहाँ, ई, ओ
- Relative(DMR)- जे, जिनका, से, जाहि
- Wh-word(DMQ)- के, की
- Indefinite(DMI)- केओ, किछु, कोनो

VERB

- It has two types -
  Main (VM) - रौपेत, खाइत, सुतैत
  Auxilliary (VAUX) - अछि, छल, थिक

- Further sub type, i.e. finite and non-finite is not reflected in Maithili at the lexical level. It is only at Local Word Grouping that Maithili reflects finiteness of its verbs.

- Therefore, the Maithili tag set does not have the second level of verb hierarchy.

ADJECTIVE

Adjective(JJ)- मोटका,मोटकी,नीक

- The adjectival participles are included within verb category (later to be under non-finite verbs).

ADVERB

- This includes only manner adverbs, adverbs of time and space are included in the Nloc (NST) under the category Noun.
- Adverb(RB)- भने, अनायास, एकाएक, फेर

POSTPOSITION (PSP)

- Postposition(PSP)-सँ, कैं, मे, लेल, दिस, धरि,पर, तर, हेतु

CONJUNCTION (CC)

- Conjunction have been taken as a top level category with two types -

  Co-ordinator(CCD) – आओर, परञ्च, मुदा, वा, आ, किंवा

  Subordintor(CCS) –जँ, तँ, कि, कियाकतँ , जखनकि

PARTICLE (RP)

Particles has five types in the Maithili tag set -

- Default(RPD)- अओ, भरि, ही, यौ, रौ,हओ,सेहो,केवल,
- Classifier(CL)- टा,गोट,गो,ठो
- Interjection(INJ)-अरे, हा, ओ, ओह
- Intensifier(INTF)-बहुत, बेसी, बेहद
- Negation(NEG)-नहि, न, जुनि

QUANTIFIERS (QT)

- General (QTF)- थोड़, बहुत, किछु, एतबा, बेसी, जतेक, ततेक, अनेक, तत्
- Cardinals (QTC)-एक, एकटा, दुई, बीसगोट, तीन, चारि
- Ordinals (QTO) -पहिल, दोसर, तेसर, चारिम

RESIDUALS (RD)

- Foreign word(RDF)-A word written in script other than the script of the original text. For example, John/RDF
- Symbol(SYM)- $, &, *,
- Punctuation (PUNC)-. , l : ;
- Unknown (UNK)- This includes those words which can not be tagged in any of the categories/types.
- Echo-words (ECH)- पानि/N_NN तानी/RD_ECH, रस्ता /N_NN तस्ता /RD_ECH,

CONCLUSION

- Till now we have tagged only 20k Maithili corpus. Further issues will be encountered after we annotate more data.

- At this stage, the present tag set provides approximately 90% coverage of the Maithili grammatical categories.

THANK YOU