Challenges and Issues in Malayalam Parts of Speech Tagging

Jisha P Jayan, Remya L B, Rajeev R R, Elizabeth Sherly
IIITM-K, Trivandrum

Overview

- What is Part Of Speech Tagging?
- What is Tagset?
- Challenges & Issues
- Conclusion
- References

What is POS Tagging?

- Part-of-speech (POS) tagging ,also known as grammatical tagging, is the process of marking the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context (i.e., relationship with adjacent and related words in a phrase, sentence, or paragraph).
- POS Tagging is the most common form of Corpus Annotation

POS Tagging

- First step in parsing
- More tractable than full parsing, intermediate representation
- Useful as a step for several other, more complex NLP tasks, e.g.
  - o Information extraction
  - o Word sense disambiguation
  - o Speech Synthesis
- Oldest task in Statistical NLP
- Easy to evaluate
- Inherently sequential

What is a Tagset ?

Tagset are names given to a set of tags from which tags are to be given to the input words in a text.

This tagset consists of grammatical tags, which may include the morphological, morpho-syntactic, semantic, and discourse level of tags.

Flat Tagset

- Flat tagset gives only the first level of information which cannot give higher level of granularity without a very large list of independent labels.
- These are easy to process as there are only a few labels.
- Flat tagset are chosen on the basis of requirement of the language.
- Eg: IIITH Tagset

IIITH Tagset

Main Tags	Representation	Main Tags	Representation
Noun	NN	Particles	RP
Noun Location	NST	Adjectives	JJ
Proper Noun	NNP	Auxillary Verb	VAUX
Pronoun	PRP	Negation	NEG
Compound Words	XC	Quantifiers	QF
Demonstration	DEM	Cardinal	QC
Post Position	PSP	Ordinal	QO
Conjuncts	CC	Question Words	WQ
Verb	VM	Intensifiers	INTF
Adverb	RB	Interjection	INJ
Symbol	SYM	Reduplication	RDP
		Unknown Words	UNK

Hierarchical Tagset

- Hierarchical tagset on the other hand uses grammatical categories and their sub categories along with other morpho-syntactic attributes.
- These are structured relative to one other, instead of using large number of independent labels and contains a small number of categories at the top level, with a number of sub categories in the form of a tree.
- Eg Bureau of Indian Standard (BIS) tagset

BIS Tagset

Sl. No	Category			Label	Annotation
	Top level	Subtype (level 1)	Subtype 2		
1	**Noun**			**N**	**N**
1.1		Common		NN	N__NN
1.2		Proper		NNP	N__NNP
1.3		Nloc		NST	N__NST
2	**Pronoun**			**PR**	**PR**
2.1		Personal		PRP	PR__PRP
2.2		Reflexive		PRF	PR__PRF
2.3		Relative		PRL	PR__PRL
2.4		Reciprocal		PRC	PR__PRC
2.5		Wh-word		PRQ	PR__PRQ
3	**Demonstrative**			**DM**	**DM**
3.1		Deictic		DMD	DM__DMD
3.2		Relative		DMR	DM__DMR
3.3		Wh-word		DMQ	DM__DMQ

BIS Tagset Contd….

4	Verb			**V**	**V**
4.1		Main		VM	V__VM
4.1.1			Finite	VF	V__VM__VF
4.1.2			Non-finite	VNF	V__VM__VNF
4.1.3			Infinitive	VINF	V__VM__VINF
4.2		Verbal		VN	V__VN
4.3		Auxiliary		VAUX	V_VAUX
5	Adjective			**JJ**	
6	Adverb			**RB**	
7	Postposition			**PSP**	
8	Conjunction			**CC**	**CC**
8.1		Co-ordinator		CCD	CC__CCD

BIS Tagset Contd….

10	**Quantifiers**			**QT**	**QT**
10.1		General		QTF	QT__QTF
10.2		Cardinals		QTC	QT__QTC
10.3		Ordinals		QTO	QT__QTO
11	**Residuals**			**RD**	**RD**
11.1		Foreign word		RDF	RD__RDF
11.2		Symbol		SYM	RD__SYM
11.3		Punctuation		PUNC	RD__PUNC
11.4		Unknown		UNK	RD__UNK
11.5		Echowords		ECH	RD__ECH

Challenges and Issues

Malayalam words are formed with combinations like noun and verb, verb, adjectives, adverbs with connectives.

It is very difficult to determine such combinations of words with other parts of speech for the categorization of that particular word.

Issues with Connective - “uM”

അവളും അവനും വരും .   { pronoun }

രാമനും സീതയും ചേർന്ന് പഠിക്കുന്നു .   { noun }

മീനു ആടുകയും , പാടുകയും ചെയ്യുന്നു .   {verb}

രാമുവിന്റേയും , രാഗുവിന്റേയും കൂടെ ഞാനും വരാം .   {psp}

ഉയരവും ,വണ്ണവും ഉള്ള മരം .   {adjective}

സീത ഇവിടേയും , അവിടേയും പോയി .   {adverb}

Issues with Connectives “oo”

അവളോ ,അവനോ ഇന്നു വരും .

രാജു ഓടിയോ , നടന്നോ വരും . {verb}

അവന്റെ അല്ലെങ്കിൽ അവളുടെ കൂടെ നിനക്ക് പോകാം .

Issues with Negations

അവൻ വരില്ല

അഴകില്ലാത്ത കുട്ടി.

അവൻ അരുതാത്തത് ചെയ്തു .

Issues with Noun - Verb Combinations

രാമനെത്തി

രാമനോടി

അവനെത്താറായോ

അവനായിട്ട്ചെയ്തു .

അവനോടായിട്ടാണ് പറഞ്ഞത് .

തിരുവനന്തപുരത്തെത്തി

Issues on Pronouns with other POS

അവൾക്കുവേണ്ടി { prp + psp }

അവനായിട്ട് { prp + verb }

അവർക്കില്ല { prp + neg }

Challenges with Avvyas

മിക്കവാറും

മുതലായവ

എന്തുകൊണ്‌ടെന്നാൽ

എങ്ങനെയെന്നാൽ

Challenges with Multiple Word Combinations

വീടിനടുത്താണ്

{ noun + noun locative + auxillary word }

പേരെടുത്തസ്ഥാനമാക്കുന്നത്

{ common noun + non finite verb + common noun + verb nonfinite}

ജലത്തിൽനിന്നുയർന്നുവന്നപ്പോൾ

{ common noun + post position + verb nonfinite+ verb nonfinite + adverb}

Conclusions

In this paper we have presented some of the issues on Malayalam language tagging while using the common Tagset for Indian languages. It is found that hierarchical tagset is more suitable for Malayalam because it keeps some morpho-syntactic features at POS level. In the present day NLP research, large volume of annotated corpora plays a significant role and is the basic building block for constructing statistical models for automatic processing of natural languages, Therefore, a widely accepted tagset for Dravidian Languages provides better results in tagging and annotation.

Applications of Tagging

- Information retrieval
- Text to speech
- Information extraction
- Linguistic research for corpora
- Higher level NLP tasks

  * Parsing

  * Semantics

  * Machine translation

References

- IIIT-Tagset. A Parts-of-Speech tag set for Indian Languages. http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf
- Bharati, Akshar, e.al.. 2006. Anncorra: Annotating Corpora. Language Technologies Research Centre, IIIT, Hyderabad
- Baskaran, S. et.al.. 2008. A Common Parts-of-Speech Tag set Framework for Indian Languages. In Nicoletta Calzolari et.al. (Eds.) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

Thank You