

# **ICON10 NLP TOOLS CONTEST: INDIAN LANGUAGE DEPENDENCY PARSING**

Kharagpur, India

Samar Husain

# Background

- Parsing morphologically rich free word order languages challenging
  - Indian Languages (IL)
  - Czech, Hebrew, Arabic, Turkish, etc.
- ICON09 tools contest on IL dependency parsing
- NAACL10 workshop on Statistical Parsing of Morphologically Rich Languages
- Non-configurationality
- Free word order
- Distributed syntactic cues
- ...
  - Large gap between labeled attachment and unlabeled attachment

# Data

- Dependency annotation follows Computational Paninian Grammar

Type	Lang.	Sent Count	Word Count	Avg. sent length
	Hindi	2,972	64632	22.69
Train	Telugu	1,400	7602	5.43
	Bangla	980	10305	10.52
	Hindi	543	12617	23.28
Devel	Telugu	150	839	5.59
	Bangla	150	1196	7.97
	Hindi	320	6589	20.59
Test	Telugu	150	836	5.57
	Bangla	150	1350	9.0

# Information in the released data

- Morphological information
- Part of Speech (POS) tag
- Chunk boundary and chunk tag
- Chunk head information
- *Vibhakti* of the head
- Dependency relation

# Information in the released data

Lang	POS	Ch	Dep	Mo	Head	Vib.
Hin	Man	Man	Man/ Auto	Man	Auto	Auto
Tel	Man	Man	Man	Auto	Auto	Auto
Ban	Man	Man	Man	Auto	Auto	Auto

Table 2. Lang: Language, Hin: Hindi, Tel: Telugu, Ban: Bangla POS: POS tags, Ch: Chunk boundaries and tags, Dep: Dependency relations, Mo: Morphological features, Head: Chunk head information, Vib: *vibhakti* of head. Man: Manual, Auto: Automatic

- For Bangla and Telugu
  - Relations between chunk heads
- For Hindi
  - Relations between words
    - An automatic tool gives intra-chunk relations (96% accurate)
- Coarse-grained and fine-grained dependency tagset

# Released data format

- Shakti Standard Format (SSF)
  - Actual annotation format
- CoNLL-X
  - Automatically converted from SSF

# SSF Representation

*raama PZala KAwa hE*

‘Ram’ ‘fruit’ ‘eat’ PRES

‘Ram eats a fruit’

1	((	NP	<fs af=‘rAma,n,m,s,3,0,’ drel=‘k1:VGF’ name=‘NP’>
1.1	rAma	NNP	<fs af=‘rAma,n,m,s,3,0,’>
	)		
2	((	NP	<fs af=‘PZala,n,m,s,3,0,’ drel=‘k2:VGF’ name=‘NP2’>
2.1	PZala	NN	<fs af=‘PZala,n,m,s,3,0,’>
	)		
3	((	VGF	<fs af=‘KA,v,m,s,,3,wA-hE,’ name=‘VGF’>
3.1	KAwa	VM	<fs af=‘KA,v,m,s,,3,wA,’>
3.2	hE	VAUX	<fs af=‘hE,v,m,s,,3,0,’>
	)		

# CoNLL Representation

*raama PZala KAwa hE*

‘Ram’ ‘fruit’ ‘eat’ PRES

‘Ram eats a fruit’

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	PHEAD	PDEPREL
1	rAma	rAma	NP	NNP	m s 3 0	3	k1	–	–
2	PZala	PZala	NP	NN	m s 3 0	3	k2	–	–
3	KAwa	KA	VP	VM	m s 3 wA	0	main	–	–
4	hE	hE	VP	VAUX	m s 3  wA-hE	3	lwg__vaux	-	-

# Evaluation measure

- Unlabeled Attachment Score (UAS),
  - the percentage of words in the sentences across the entire test data that have correct parents.
- Label Accuracy (LA),
  - the percentage of words with correct dependency label
- Labeled Attachment Score (LAS)
  - the percentage of words with correct parent and correct dependency label

# Contest

- 15 teams registered
- 6 out of 15 teams submitted the results
  - 4 systems for all the languages
  - 2 systems for one language
- August-November
- Submission of systems
  - Two separate outputs (Fine-grained and Coarse-grained)

# Results (Fine-grained tagset)

- UAS
  - Hindi: 94.78 (Attardi et al.)
  - Telugu: 91.82 (Kosaraju et al.)
  - Bangla: 87.41 (Attardi et al.)
- LS
  - Hindi: 90.00 (Kosaraju et al.)
  - Telugu: 71.95 (Kosaraju et al.)
  - Bangla: 73.47 (Attardi et al.)
- LAS
  - Hindi: 88.63 (Kosaraju et al.)
  - Telugu: 70.12 (Kosaraju et al.)
  - Bangla: 70.66 (Attardi et al.)

# Overall results

1. Kosaraju et al.
2. Kolachina et al.

# Programme Schedule

## (16:00 – 17:30)

- **Bidirectional Dependency Parser for Indian Languages.**  
Aswarth Abhilash and Prashanth Mannem
- **Dependency Parsing of Indian Languages with DeSR.**  
Giuseppe Attardi, Stefano Dei Rossi and Maria Simi
- **A Two Stage Constraint Based Hybrid Dependency Parser for Telugu.**  
Sruthilaya Reddy Kesidi, Prudhvi Kosaraju, Meher Vijay and Samar Husain
- **Experiments with MaltParser for parsing Indian Languages.**  
Sudheer Kolachina, Prasanth Kolachina, Manish Agarwal and Samar Husain
- **Experiments on Indian Language Dependency Parsing.**  
Prudhvi Kosaraju, Sruthilaya Reddy Kesidi, Vinay Bhargav Reddy Ainavolu and Puneeth Kukkadapu

Thanks!!