

ICON₀₉ NLP TOOLS CONTEST: INDIAN LANGUAGE DEPENDENCY PARSING

Samar Husain
LTRC, IIIT-Hyderabad.

Outline

- Introduction
- Background
 - Task
 - IL Treebanks
 - Evaluation metric
- Contest
- Results

Introduction

- Broad coverage parser for IL
 - Very crucial
 - Required as a preprocessing step for almost all NLP applications:
 - MT systems,
 - IE,
 - co-reference resolution,
 - NLU, etc.

Pre-parsing

- Levels of analysis before parsing
 - Morphological analysis
 - Shallow parsing
- We parse after the above processing is done

Example

- rAma ne mohana ko puswaka xI |
‘Ram’ ‘ERG’ ‘Mohana’ ‘DAT’ ‘book’ ‘gave’
Ram gave a book to Mohana.

Example – POS tagged and Chunked (SSF Format)

```
1      ((      NP
1.1    rAma    NNP
1.2    ne      PREP
      ))
2      ((      NP
2.1    mohana  NNP
2.2    ko      PREP
      ))
3      ((      NP
3.1    puswaka NN
      ))
4      ((      VG
4.1    xI      VEM
4.2    |       SYM
      ))
```

Example – Morph Analysis

1	((NP	<name=1>
1.1	rAma	NNP	<af=rAma,n,m,s,,1,ne,>
)		
2	((NP	<name=2>
2.1	mohana	NNP	<af=mohana,n,m,s,,1,ko,>
)		
3	((NP	<name=3>
3.1	puswaka	NN	<af=puswaka,n,f,s,,0,0,>
)		
4	((VG	<name=4>
4.1	xI	VFM	<af=xe,v,f,s,any,,,yA>
4.2		SYM	
)		

xe,	v,	f,	s,	any,	,	,	yA
lex	cat	gen	num	per	cas	vib	tam
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)

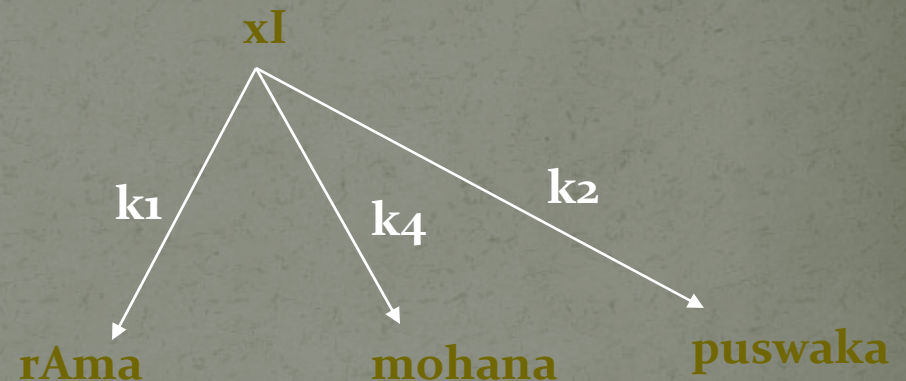
Example – Parsed Output (SSF Format)

```
1      ((      NP      <af=rAma,n,m,s,,1,ne,/drel=varg__k1:4/name=1>
1.1    rAma    NNP     <af=rAma,n,m,s,,1,ne,>
      ))
2      ((      NP      <af=mohana,n,m,s,,1,ko,/drel=varg__k4:4/name=2>
2.1    mohana NN      <af=mohana,n,m,s,,1,ko,>
      ))
3      ((      NP      <af=puswaka,n,f,s,,0,0,/drel=varg__k2:4/name=3>
3.1    puswaka NN     <af=puswaka,n,f,s,,0,0,>
      ))
4      ((      VG      <af=xe,v,f,s,any,,,yA/name=4>
4.1    xI      VFM     <af=xe,v,f,s,any,,,yA>
4.2    |      SYM
      ))
```

Example – Parsed Output

$X = w_0 w_1 w_2 \dots w_n; w_0 = \text{root}$
 $L = \{l_1, l_2, \dots, l_{|L|}\}$
 $G = (V, A)$
 $V = \{0, 1, \dots, n\}$ is a vertex set
 A is the arc set; $(i, j, k) \in A$

G is a tree



IL Treebanks

- Languages
 - Telugu
 - Bangla
 - Hindi
- Training Data

Language	No. of Sentences	Word Count	Average sentence length
Telugu	1,400	7602	5.43 words
Bangla	980	10305	10.52 words
Hindi	1,500	28522	19.01 words

- Testing and development
 - 150 for all three

- Based on the Paninian grammar (Bharati et al., 1995, Begum et al., 2008)
 - dependency trees
 - syntactico-semantic labels called *karakas*
- Released data also contained automatically computed morphological, head information, etc.

Evaluation metrics

- CoNLL dependency parsing shared task 2008 (Nivre et al., 2008)
- UAS: Unlabeled attachment accuracy
- LAS: Label attachment accuracy
- LA: Label accuracy

Contest

- 1st August 09 – 1st December 09
 - <http://ltrc.iiit.ac.in/nlptools2009>
- 8 teams
- Evaluation done over 2 rounds

- 1st round contained fine-grained tags
 - 28
- 2nd round contained coarse-grained tags
 - 12

- Same dataset for both the rounds

Tagset: 1st Round

- Fine-grained
- Total: 28

k1	karta (doer/agent/subject)
k2	karma (object/patient)
k3	karana (instrument)
k4	sampradaana (recipient)
k5	apaadaana (source)
k7t	kaalaadhikarana (location in time)
k7p	deshadhikarana (location in space)
k7	vishayaadhikarana (location elsewhere)
ras	upapada__ sahakaarakatwa (associative)
rd	prati upapada (direction)
rh	hetu (cause-effect)
k*u	saadrishya (similarity)
rt	taadarthya (purpose)
k1s	vidheya karta (karta samanadhikarana)
k2s	vidheya karma (karma samanadhikarana)
r6	shashthi (possessive)
pk1	prayojaka karta (causer)
mk1	madhyastha karta (causer2)
jk1	prayojya karta (causee)
nmod	Noun modifiers
jjmod	Adjectival modifiers
adv	kriyaavisheshana ('manner adverbs' only)
rad	Address words
ccof	Conjunct of relation
pof	Part of relation
nmod_relc	Noun modifier of the type relative clause
jjmod_relc	Adjectival modifier of the type relative clause.
rbmod_relc	Adverbial modifier of the type relative clause

Tagset: 2nd Round

- Coarse-grained
- Total: 12
 1. k1-ext
 2. k2-ext
 3. k7-ext
 4. vmod-rest
 5. nmod
 6. r6
 7. relc
 8. rbmod
 9. jjmod
 10. ccof
 11. fragof
 12. main

Parsing methods used by participants

- Transition based methods: Malt
- Constraint based methods
- Bidirectional parsing
- Hybrid methods
 - Malt + postprocessing
 - Malt + MST + DZ
 - Constraints + ML
 - CRF + postprocessing

Results

- Average score over 2 rounds

Languages	UAS	LAS	LS
Telugu	84.78	59.73	61.41
Bangla	81.96	66.97	71.1
Hindi	88.78	72.83	75.59

Results

- Best performing systems

Languages (Team)	UAS	LAS	LS
Telugu (Mannem)	85.76	65.01	66.21
Bangla (De et al.)	90.32	84.29	85.85
Hindi (Ambati et al.)	90.22	79.33	81.66

- Mannem: Bidirectional parsing
- De et al.: Constraint based parsing
- Ambati et al.: Malt

ICON Tools Contest Final Round Results

LAS

Groups	Teams	University	Hindi-fine	Bangla-fine	Telugu-fine	Ave-fine	Hindi-coarse	Bangla-coarse	Telugu-coarse	Ave-coarse	Ave-Final	POSITION
1	Ghosh	JU, India		53.9		53.9		19.04			53.9	
2	Nivre	UU, Sweden	73.36	70.45	57.63	67.15	78.2	76.07	62.44	72.24	69.69	2 nd
3	Zeman	CU, Czech	68.25	66.6	50.94	61.93	73.88	71.49	56.43	67.27	64.6	4 th
4	Hasan	UTD, USA										
5	De et al.	ISI, Kolkata		79.81		79.81		84.29		84.29	82.05	
6	Ambati et al.	IIT-H, India	74.48	72.63	60.55	69.22	79.33	78.25	65.01	74.2	71.71	1 st
7	Chatterji et al.	IIT-kgp, India	61.59	60.46		61.03					61.03	
8	Vijay et al.	IIT-H, India	62.2			62.2					62.2	
9	Mannem	IIT-H, India	71.63	67.74	59.86	66.41	76.9	70.34	65.01	70.75	68.58	3 rd
Average			68.59	67.37	57.25	65.21	77.08	66.58	62.22	73.75	66.72	

UAS

			Hindi-fine	Bangla-fine	Telugu-fine	Ave-fine	Hindi-coarse	Bangla-coarse	Telugu-coarse	Ave-coarse	Ave-Final	
1	Ghosh	JU, India		74.09		74.09		32.88			74.09	
2	Nivre	UU, Sweden	89.79	88.66	84.73	87.73	89.36	88.97	86.28	88.2	87.97	2 nd
3	Zeman	CU, Czech	88.32	86.68	82.5	85.83	88.49	86.89	81.3	85.56	85.7	4 th
4	Hasan	UTD, USA										
5	De et al.	ISI, Kolkata		90.32		90.32		90.32		90.32	90.32	
6	Ambati et al.	IIT-H, India	90.14	88.45	86.28	88.29	90.22	90.22	85.25	88.56	88.43	1 st
7	Chatterji et al.	IIT-kgp, India	84.95	82.21		83.58					83.58	
8	Vijay et al.	IIT-H, India	85.55			85.55					85.55	
9	Mannem	IIT-H, India	88.24	85.33	86.11	86.56	88.06	83.56	85.76	85.79	86.18	3 rd
Average			87.83	85.11	84.91	85.24	89.03	78.81	84.65	87.69	85.23	

LS

			Hindi-fine	Bangla-fine	Telugu-fine	Ave-fine	Hindi-coarse	Bangla-coarse	Telugu-coarse	Ave-coarse	Ave-Final	
1	Ghosh	JU, India		61.71		61.71		29.14			61.71	
2	Nivre	UU, Sweden	75.26	72.94	58.49	68.9	81.06	79.6	62.95	74.54	71.72	2 nd
3	Zeman	CU, Czech	72.66	71.28	54.2	66.05	77.94	76.59	60.89	71.81	68.93	4 th
4	Hasan	UTD, USA										
5	De et al.	ISI, Kolkata		81.27		81.27		85.95		85.95	83.61	
6	Ambati et al.	IIT-H, India	76.38	75.34	61.58	71.1	81.66	81.69	66.21	76.52	73.81	1 st
7	Chatterji et al.	IIT-kgp, India	63.32	65.97		64.65					64.65	
8	Vijay et al.	IIT-H, India	65.88			65.88					65.88	
9	Mannem	IIT-H, India	73.7	69.93	60.72	68.12	79.24	73.05	66.21	72.83	70.48	3 rd
Average			71.2	71.21	58.75	68.46	79.98	71	64.07	76.33	70.1	

References

- Rafiya Begum, Samar Husain, Arun Dhvaj, Dipti Misra Sharma, Lakshmi Bai and Rajeev Sangal. 2008. *Dependency Annotation Scheme for Indian Languages*. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP)*. Hyderabad, India. 2008.
- Bharati, A., V. Chaitanya and R. Sangal. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi
- J. Nivre and J. Hall and S. Kubler and R. McDonald and J. Nilsson and S. Riedel and D. Yuret. 2007b. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

Thank you

