

Dependency Parsers for Indian Languages

Samar Husain

Language Technologies Research Centre
IIIT-Hyderabad, India.

samar@mail.iiit.ac.in

Due to the availability of annotated corpora for various languages since the past decade, data driven parsing has proved to be immensely successful. Unlike English, however, most of the parsers for morphologically rich free word order (MoR-FWO) languages (such as Czech, Turkish, Hindi, etc.) have adopted the dependency grammatical framework. It is well known that for MoR-FWO languages, dependency framework provides ease of linguistic analysis and is much better suited to account for their various structures (Shieber, 1975; Mel'cuk, 1988; Bharati et al., 1995).

The NLP tools contests are regular events held as part of the International Conference on Natural Language Processing (ICON) and cater to various NLP tasks. This year, the contest focused on Indian language dependency parsing. Three languages, namely Telugu, Bangla and Hindi were explored. 8 teams participated in the event.

Manually annotated data in all the three languages was given to the participating teams. The data was POS tagged, chunked and marked for dependency information. The dependency relations are based on the Paninian grammatical model (Bharati et al., 1995; Begum et al., 2008). The data also contained automatically computed morphological, head information etc. General statistics for the released training data is shown in table 1.

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

Table 1.

The development and the testing set for all the three languages had 150 sentences each. The released annotated data, although small, can provide considerable insight into various parsing issues. The contest is intended as a first step in this direction. As the results suggest, fairly high accuracies have been obtained in spite of the small data size.

The teams submitted the results in 2 rounds. In the first round the dependency tagset was fine-grained and was larger than the tagset used in the 2nd round. For both the rounds the data used was identical. The performance of the system was measured in terms of standard measures such as Unlabelled attachment score (UAS) and Labelled attachment score (LAS) (Nivre et al., 2007b). The average scores over the 2 rounds for different languages are mentioned in table 2.

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

Table 2. Average scores

Table 3 shows the results of the best performing system in the three languages.

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

Table 3. Best results on course grained tagset

Many methods were explored. Ghosh et al. (2009) use a CRF based hybrid method. Nivre (2009), Ambati et al. (2009) and Chatterji et al. (2009) used the well known transition based dependency parsing (Nivre et al., 2007a). Zeman (2009) combines various well known dependency parsers forming a superparser by using a voting method. De et al. (2009), Yeleti and Deepak (2009) use a constraint based approach. Mannem (2009) tried bi-directional incremental parsing and perceptron learning to attack the problem.

Teams that submitted results on all the three languages were considered for ranking. De et al. (2009) got best results for Bangla; however since they participated only for Bangla they have not been ranked. Consolidated result of all the systems for both the rounds can be seen in Figure 1.

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.	IIIT-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.	IIIT-kgp, India	61.59	60.46		61.03						61.03
8	Vijay et al.	IIIT-H, India	62.2			62.2						62.2
9	Mannem	IIIT-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.	IIIT-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.	IIIT-kgp, India	84.95	82.21		83.58						83.58
8	Vijay et al.	IIIT-H, India	85.55			85.55						85.55
9	Mannem	IIIT-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.	IIIT-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.	IIIT-kgp, India	63.32	65.97		64.65						64.65
8	Vijay et al.	IIIT-H, India	65.88			65.88						65.88
9	Mannem	IIIT-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

Figure 1. Consolidated results

References

B. R. Ambati, P. Gadde and K. Jindal. 2009. Experiments in Indian Language Dependency Parsing. *In Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing. Hyderabad, India. 2009.*

- R. Begum, S. Husain, A. Dhawaj, D. M. Sharma, L. Bai, and R. Sangal. 2008. Dependency annotation scheme for Indian languages. In *Proceedings of IJCNLP-2008*.
http://www.iiit.net/techreports/2007_78.pdf
- A. Bharati, V. Chaitanya and R. Sangal. 1995. *Natural Language Processing: A Paninian Perspective*, Prentice-Hall of India, New Delhi, pp. 65-106. lrc.iiit.ac.in/downloads/nlpbook/nlp-panini.pdf
- S. Chatterji, P. Sonare, S. Sarkar and D. Roy. 2009. Grammar Driven Rules for Hybrid Bengali Dependency Parsing. In *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India. 2009*.
- S. De, A. Dhar, and U. Garain. 2009. Structure Simplification and Demand Satisfaction Approach to Dependency Parsing in Bangla. In *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India. 2009*.
- A. Ghosh, P. Bhaskar, A. Das, and S. Bandyopadhyay. 2009. Dependency Parser for Bengali: the JU System at ICON 2009. In *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India. 2009*.
- P. Mannem. 2009. Bidirectional Dependency Parser for Hindi, Telugu and Bangla. In *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India. 2009*.
- I. A. Mel'cuk. 1988. *Dependency Syntax: Theory and Practice*, State University Press of New York.
- J. Nivre. 2009. Parsing Indian Languages with MaltParser. In *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India. 2009*.
- J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov and E Marsi. 2007a. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2), 95-135.
- 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*.
- S. M. Shieber. 1985. Evidence against the context-freeness of natural language. In *Linguistics and Philosophy*, p. 8, 334–343.
- M. V. Yeleti and K. Deepak. 2009. Constraint based Hindi dependency parsing. In *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India. 2009*.
- D. Zeman. 2009. Maximum Spanning Malt: Hiring World's Leading Dependency Parsers to Plant Indian Trees. In *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India. 2009*.