Partial Parsing as a Method to Expedite Dependency Annotation of a Hindi Treebank

Mridul Gupta, Vineet Yadav, Samar Husain and Dipti Misra Sharma
Language Technologies Research Center,
International Institute of Information Technology, Hyderabad
INDIA – 500 032
E-mail: mridulgupta@students.iiit.ac.in, vineetyadav@students.iiit.ac.in, samar@mail.iiit.ac.in,
dipti@mail.iiit.ac.in

Abstract

The paper describes an approach to expedite the process of manual annotation of a Hindi dependency treebank which is currently under development. We propose a way by which consistency among a set of manual annotators could be improved. Furthermore, we show that our setup can also prove useful for evaluating when an inexperienced annotator is ready to start participating in the production of the treebank. We test our approach on sample sets of data obtained from an ongoing work on creation of this treebank. The results asserting our proposal are reported in this paper. We report results from a semi-automated approach of dependency annotation experiment. We find out the rate of agreement between annotators using Cohen's Kappa. We also compare results with respect to the total time taken to annotate sample data-sets using a completely manual approach as opposed to a semi-automated approach. It is observed from the results that this semi-automated approach when carried out with experienced and trained human annotators improves the overall quality of treebank annotation and also speeds up the process.

1. Introduction

One of the most important linguistic resources is a treebank. Resources such as these are of immense utility to various NLP tasks such as syntactic parsing, natural language understanding, MT etc. and have been endorsed universally. Also, treebanks are important for machine learning and for extracting out various kinds of linguistic information. Treebank annotation is carried out to encode linguistic information at different levels such as morphological, syntactic, syntactico-semantic, and discourse. Consistency and quality are important aspects during treebank annotation. Much focus has been given to annotating syntactic structures using various linguistic paradigms during the last decade. This is understandable as building efficient syntactic tools, such as parsers, is very crucial for various NLP applications. In this paper our focus would be syntactico-semantic dependency annotation. The dependency annotation scheme followed is based on computational Paninian grammar (CPG) (Bharati et al., 1995).

In this paper, we further elaborate on a method proposed by Bharati et al., (2009a), which can lead to faster annotation of sentences in Hindi, without compromising accuracy and consistency. Bharati et al. propose an automatic annotator tool, 'simple parser for Hindi'. The tool uses a knowledge-based methodology which relies on exploiting certain linguistic and syntactic cues deduced from the manually annotated training data (development data) containing about 1800 sentences in Hindi. In this paper we use their tool's output as the data over which manual annotation is done. Use of such a partially parsed output is justified as it involves less analysis as compared to that for a fully generated parse (Bharati et al., 2009a).

Indian languages, like Hindi, are relatively rich in morphology and have relative free word-order. Words are grouped into chunks (Bharati et al., 2007). These words within the chunks are fixed and cannot move. These chunks can move around within the sentence which accounts for its free-word order. Wherein, the scope of partial parsing is to capture dependency relations between the chunks only. Hence, it is important that its methodology (data-driven or knowledge-based) is independent of the word-order and derives cues from the morphological features of words. By partial parse, we mean a parsed output which has some important dependency labels annotated between chunks using a set of simple yet effective rules.

This approach, as proposed in (Bharati et al., 2009a), with the use of an automated tool (a) can help in the process of faster and equally (if not more) accurate annotation as opposed to all the annotation done manually, and (b) could be useful for inexperienced annotators to learn the process of dependency annotation over a certain number of iterations. The paper describes an approach to expedite the process of manual annotation of a Hindi dependency treebank. We propose a way by which consistency among a set of manual annotators could be improved. Furthermore, we show that our setup can also prove useful for evaluating when an inexperienced annotator is ready to start participating in the production of the treebank. We test our approach on sample sets of data obtained from an ongoing work on creation of this treebank. The results asserting our proposal are reported in this paper.

The paper is divided as follows. Section 1 talked about the introduction. Section 2 is an overview of some of the works related to this paper. In section 3, we provide a brief description of the Hindi treebank. Section 4 describes the experiments conducted. Results and observations from those experiments follow in section 5. Section 6, finally concludes the paper.

2. Related Work

This semi-automated approach for annotation holds merit and has previously been used for annotation of treebanks. Such a methodology has been followed for annotation of treebanks in languages such as German (Brants and Skut, 1998), English on the PARC 700 Dependency Treebank (King et al., 2003) etc.

Inter-annotator agreement is extensively used for evaluation of consistency of annotation of corpora. The inter-annotator agreement to evaluate quality has previously been used extensively on resources like Penn Treebank (Marcus et al., 1993), NEGRA corpus (Brants, 2000), Brown corpus (Francis and Kucera 1982), in Penn Discourse Treebank Annotation (Miltsakaki et al., 2004a; Miltsakaki et al., 2004b), Arabic treebank (Habash and Roth, 2009) and others. Apart from inter-annotator agreement, Miltsakaki et. al (2004a) also used rate of disagreement for analysis of the Penn discourse Treebank. In (Marsi and Krahmer, 2005) agreement rate was used to evaluate the system on an aligned parallel Dutch corpus.

3. The Hindi Dependency Treebank (HDT)

A multi-layered and multi-representational treebank for Hindi (Bhatt et al., 2009; Xia et al., 2009) is being developed. The treebank will have dependency, verb-argument (PropBank, Palmer et al., 2005) and phrase structure (PS) representation. Automatic conversion from dependency structure (DS) to phrase structure, (PS) is planned. However, the focus of the current paper is to ascertain the effectiveness of partial dependency parsing in helping the manual annotators to expedite the process of dependency annotation of the treebank. The dependency treebank contains information encoded at the morpho-syntactic (morphological, part-of-speech and chunk information) and syntactico-semantic (dependency) levels. Each sentence is represented in SSF format (Bharati et al., 2007). POS and chunk information is encoded following a set of guidelines (Bharati et al., 2006). The guidelines for the dependency framework (Bharati et al., 2009b) have been adapted from the Paninian grammar (Bharati et al., 1995). For Indian languages, like Hindi, Paninian dependency scheme has been shown to be effective in (Begum et al., 2008).

4. Procedure and Experimental Setup

To evaluate the effectiveness of our approach we used some sample sets of data for testing. The sample sets were obtained from a previously validated gold standard Hindi data of about 50,000 tokens taken from HDT (see Section 3). Three sample sets were used for performing the experiment, labeled as S-1, S-2 and S-3. Each set consisted of 25 sentences. Five annotators were employed to carry out the experiment. These annotators were divided into two groups, namely, A (3 annotators, A1, A2 and A3) and B (2 annotators, B1 and B2). The data given to these annotators was previously unseen to them. Annotators in group A were experienced and had been doing dependency annotation for at least 8 months. They were well-trained annotators. They worked only on sample set S-1. Group B annotators were less experienced than group A and had been doing annotation for the past two months at the time of conducting our experiments. Although, B1 was slightly more experienced than B2. Group B annotated all the three sets.

We also tested whether annotation quality of group B over the three sets improved or not. The annotators performed annotation in two different modes:

(a). For the first mode, the annotators simply annotated the sample sets of data all by themselves without using any automated tool. Time taken to annotate was also noted down.

(b). For the second mode of annotation, they annotated and corrected the output of the automatic annotator for the same sample dataset. Care was taken to ensure that there was enough gap in terms of time between the two modes of annotation. There was a time lapse of at least more than a week between the two modes of annotation. This ensured that the annotators did not remember their previous annotation values.

For each mode, inter annotator agreement was calculated to evaluate and compare the levels of consistency. Also, accuracy (precision) of annotation was evaluated by comparing it against the corresponding reference gold standard data prepared by the guideline setters.

4.1 Inter Annotator Agreement

Inter-annotator agreement among annotators is a good way to determine consistency in the process of annotation. It also helps to do error analysis in a more focused manner. Thus, we chose Cohen's measure (Kappa, κ) (Cohen, 1960) of finding out agreement between annotators, as it is a widely used measure, which also accounts for agreement by chance rather than simply calculating the percentage of agreement. Although, there has been criticism of the robustness of this method (Di Eugenio, 2000), it still provides a standardized way of estimating the agreement rate across annotators.

Cohen’s kappa agreement is given by the following equation:

$$\kappa = \frac{p(a) - p(e)}{1 - p(e)}$$

where, $p(a)$ is the observed probability of agreement between two annotators, and $p(e)$ is the theoretical probability of chance agreement, using the annotated sample of data to calculate the probabilities of each annotator.

The co-efficient of Cohen's measurement has been previously applied for evaluating consistency in works on treebanks and other corpora such as the Hinoki treebank for Japanese, (Bond et al., 2006), EPEC treebank for Basque (Uria et al., 2009), Wordnet Semcor and DSO corpora (Ng et al., 1999), spoken Danish corpus (Paggio, 2006). We use the standard metric for the interpretation of kappa values (Landis and Koch, 1977). Table 1 as shown gives a measure to interpret kappa values devised by Landis and Koch.

Kappa value	Degree of Agreement
<0	None
0 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Perfect

Table 1: Landis and Koch interpretation of Cohen's kappa.

As noted from the table above, values greater than 0.61 are considered to be very good agreement rate between any one pair of annotators.

4.2 Accuracy Measurement

For evaluating accuracy of the annotators, we used two metrics, LAA[1] and LA[2]. LAA, is the precision obtained after looking at the correctness of the attachment and the label between a child-parent pair. On the other hand, LA, is the precision score obtained looking at only the label between a child and its parent.

5. Results and Observations

We present the results in this section for the experiments conducted on the sample data sets of Hindi. We show results (L(κ) [3] and LA(κ) [4]) at annotation done in each iteration by the annotators. We also report the total time taken in annotation for each of the two modes in the subsequent figures. In Tables 3(a), (b), we show the respective accuracies (precision) recorded by the experienced annotators of group A for the two modes of annotation.

Annotators	Precision (LA)	Precision (LAA)
A1	86.4%	81.9%
A2	82.9%	78.9%
A3	82.4%	78.4%

Table 2(a): Accuracy of A1, A2 and A3 on the first sample data set (Set-1) for the first mode of annotation.

Table 2(b) shows the precision values for the annotation done by each annotator in group A when they were provided with the same dataset, now partially annotated generated by the automatic annotator.

Table 3 shows accuracy figures in terms of precision using the automatic annotator only, on the three sample sets of data.

Annotators	Precision (LA)	Precision (LAA)
A1	85.8%	81.5%
A2	83.9%	79.4%
A3	83.4%	79.4%

Table 2(b): Accuracy of A1, A2 and A3 on the first sample data set (Set-1) for the second mode of annotation.

Sample Sets	Precision (LA)	Precision (LAA)
S-1	73.4%	65.1%
S-2	72.2%	64.8%
S-3	74.2%	65.7%

Table 3: Accuracy of automatic annotator for the three sample sets of data.

From tables 2(a) and (b), it is clear that the annotation quality of group A remained, more or less, the same for both modes of annotation. The time gap between the two modes was about a month, during which they had been performing annotation on other data sets. Therefore, one can safely assume that the annotation in the two modes was completely independent.

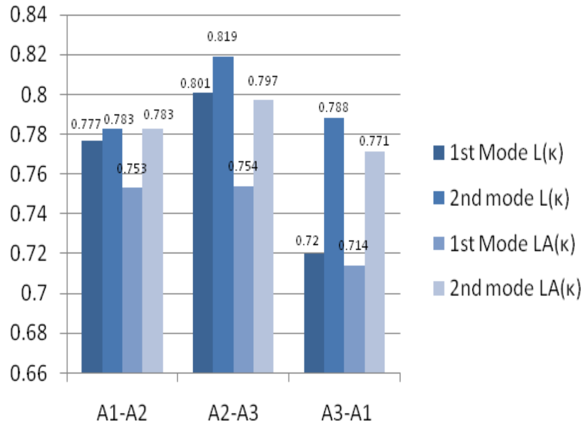

Figure 1(a): Total agreement rate at first and second modes for A1, A2 and A3.

Figure 1(a) depicts agreement rate for annotators, A1, A2, A3. It is noted that the agreement rate among them went up when the automatic annotator was used as a preprocessing tool for dependency annotation (second mode). Moreover, the time taken by them also reduced by a considerable amount (cf. figure 1(b)) in that mode. Also, tables 2(a), (b) reveal that the overall accuracy of the annotation in the second mode is at least as good as that for the first mode. But, group A annotators employed were unavailable to carry out more iterations. Thus, only one complete iteration of our experiment could be performed by them. Nonetheless, they recorded high agreement rate among themselves in the first iteration itself, which re-affirms our hypothesis of semi-automated

[1] LAA: Labeled Attachment Accuracy
[2] LA: Labeled Accuracy
[3] L(κ): Labeled Kappa
[4] LA(κ): Labeled Attachment Kappa

process being useful in the process of dependency annotation with experienced annotators.

To ascertain the validity of our claim when the same procedure is applied for relatively inexperienced annotators, we performed this experiment with another set of annotators (group B).

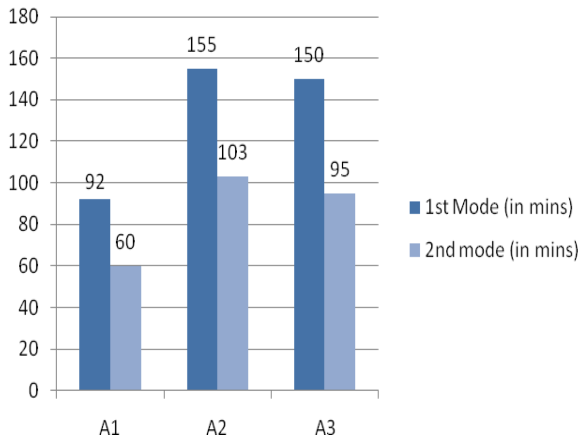

Figure 1(b): Total time taken (in mins) at each mode of annotation by A1, A2 and A3.

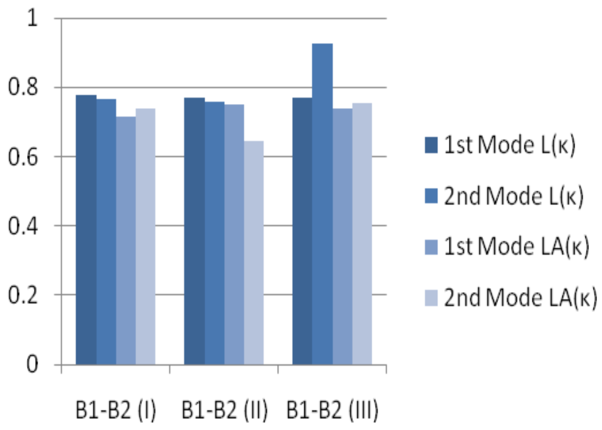

Figure 2(a): Agreement rate for B1-B2 at different iterations.

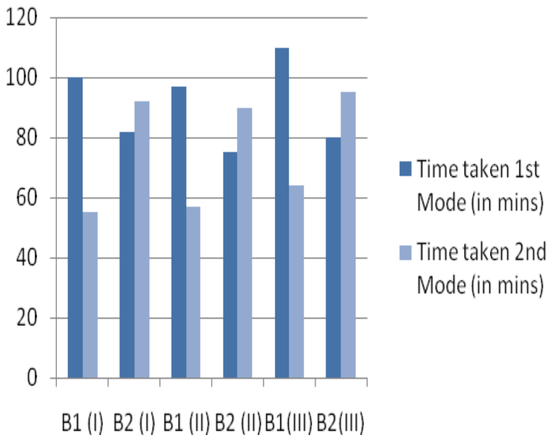

Figure 2(b): Total time taken (in mins) for each mode at each iteration.

Figures 2(a) and (b) reveal the figures for annotation consistency between the two annotators of group B, as well as the total time taken by them to annotate each sample set. In tables 4(a) and (b), accuracy numbers for B1 and B2 for the two modes at all the three iterations have been shown.

Annotators at each iteration	Precision (LA)	Precision (LAA)
B1 (1st)	78.9%	73.4%
B1 (2nd)	77.1%	71.9%
B1 (3rd)	75.3%	74.1%
B2 (1st)	81.2%	77.7%
B2 (2nd)	78.4%	74.8%
B2 (3rd)	78.3%	75.5%

Table 4(a): Accuracy of B1 and B2 for the first mode of annotation at each iteration.

Tables 4(a), (b) and figures 2(a), (b) depict numbers for annotation quality of B1 and B2.

Annotators at each iteration	Precision (LA)	Precision (LAA)
B1 (1st)	82.4%	77.4%
B1 (2nd)	78.2%	73.1%
B1 (3rd)	80.8%	78.8%
B2 (1st)	76.8%	71.7%
B2 (2nd)	76.5%	69.7%
B2 (3rd)	82.5%	76.4%

Table 4(b): Accuracy of B1 and B2 for the second mode of annotation at each iteration.

Note that B1 records greater accuracy for mode 2 at the first iteration while taking less time. On the other hand, B2, who was an even less experienced annotator, recorded lesser accuracy in the second mode as compared to the first mode for the first iteration. This may be attributed to the fact that quite a few of his/her judgments get influenced by the output of the automatic annotator. This causes confusion for the annotator. Thus, accuracy is somewhat lowered for B2. As a result of this erroneous annotation by B2, agreement rate between the two drops.

To ascertain whether the inexperienced group B annotator understood the guidelines, we used two more Hindi sample sets for annotation in two different iterations. Annotators of group A already recorded higher agreement rate and accuracy in the second mode for the first iteration.

As noted from figures 2(a), (b) and tables 4(a), (b), B1 records considerably less time for annotation and higher accuracy in second mode for all three sets. But, the same cannot be said of B2 for the first two iterations. However, as B2 got more and more familiar with the pattern of annotation, in the third and final iteration, his/her annotation quality improved. As a result, agreement rate between B1 and B2 improved in the second mode for this particular iteration.

6. Conclusion

We reported some key observations from a Hindi dependency annotation experiment conducted in two different modes. From the observations, it seems that a semi-automated process is an effective way of doing dependency annotation of treebanks when the human annotators are trained and experienced. We noted in the experiment that the time and effort of human annotators reduced. Also, an improvement was observed in accuracy and consistency among the annotators. A greater degree of consistency leads to quality assurance.

On the other hand, an inexperienced annotator can find the process pretty helpful in determining when is he/she ready for participation in the dependency annotation process. This also ascertains the fact that treebank annotation is not a trivial task and supervision is required for carrying out the task in an efficient manner.

7. Acknowledgements

We thank the reviewers for their insightful comments. We also thank the annotators for annotating the data sets used for our experiment. The work reported in this paper was supported by the NSF grant (Award Number: CNS 0751202; CFDA Number: 47.070).

8. References

Begum, R., Husain, S., Dhwaj, A., Sharma, D.M., Bai, L., Sangal, R. (2008). Dependency annotation scheme for Indian languages. In *Proceedings of IJCNLP-2008.*

Bharati, A., Chaitanya, V., Sangal, R. (1995). *Natural Language Processing: A Paninian Perspective*, Prentice-Hall of India, New Delhi, pp. 65-106.

Bharati, A., Gupta, M., Yadav, V., Gali, K., Sharma, D.M. (2009a). Simple Parser for Indian Languages in a Dependency Framework. In *Proc. of the Third Linguistic Annotation Workshop at 47th ACL and 4th IJCNLP*.

Bharati, A., Sangal, R., Sharma, D.M. (2007). SSF: Shakti Standard Format Guide. *Technical Report, TR-LTRC-33,* Language Technologies Research Centre, IIIT-Hyderabad, India.

Bharati, A., Sangal, R., Sharma, D.M., Bai, L. (2006). AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages. *Technical Report (TR-LTRC-31),* Language Technologies Research Centre, IIIT-Hyderabad, India.

Bharati, A., Sharma, D.M., Husain, S., Bai L., Begum, R., Sangal, R. (2009b). AnnCorra: TreeBanks for Indian Languages, Guidelines for Annotating Hindi TreeBank. *http://ltrc.iiit.ac.in/MachineTrans/research/tb/DS-guidelines/DS-guidelines-ver2-28-05-09.pdf*

Bhatt, R., Narasimhan, B., Palmer, M., Rambow O., Sharma, D.M., Xia. F. (2009). Multi-Representational and Multi-Layered Treebank for Hindi/Urdu. In *Proc. of the Third Linguistic Annotation Workshop at 47th ACL and 4th IJCNLP.*

Bond, F., Fujita, S., Tanaka, T. (2006). The Hinoki Syntactic and Semantic Treebank of Japanese. *Language Resources and Evaluation*, Vol. 40, No. 3/4, Asian Language Processing: State-of-the-Art Resources and Processing, pp. 253-261.

Brants, T. (2000). Inter-Annotator Agreement for a German Newspaper Corpus. In Proc. of the Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Greece.

Brants. T., Skut, W. (1998). Automation of Treebank Annotation. In *Proc. of New Methods in Language Processing (NeMLaP-98).*

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, pp. 37–46.

Di Eugenio., B. (2000). On the usage of Kappa to evaluate agreement on coding tasks. In *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC-00)*, Athens, Greece.

Francis, W.N., Kucera, H. (1982). Frequency Analysis of English Usage: lexicon and grammar. *Journal of English Linguistics*, Vol. 18, No. 1, pp. 64-70, Houghton Mifflin, Boston, MA.

Habash, N., Roth, R. (2009). CATiB: The Columbia Arabic Treebank. In *Proc. of 47th ACL- 4th IJCNLP.*

King, T.H., Crouch, R., Riezler, S., Dalrymple, M., Kaplan, R.M. (2003). The PARC 700 Dependency Bank. In *Proc. of Workshop on Linguistically Interpreted Corpora at the European Association for Computational Linguistics*.

Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics.* Vol. 33, pp. 159—174.

Marcus, M.P., Marcinkiewicz, M.A., Santorini, B. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, Volume 19, Issue 2, pp. 313 – 330.

Marsi, E., Krahmer, E. (2005). Classification of semantic relations by humans and machines. In *Proc. of the ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*.

Miltsakaki, E., Prasad, R., Joshi, A., Webber, B. (2004a). Annotating Discourse Connectives and Their Arguments. In *Proc. of the HLT/NAACL Workshop on Frontiers in Corpus Annotation.*

Miltsakaki, E., Prasad, R., Joshi, A., Webber, B. (2004b). The Penn Discourse TreeBank. In *Proc. of the Language Resources and Evaluation Conference, Portugal*.

Ng, H. T., Lim, D.C.Y., Foo, S.K. (1999). A Case Study On Inter-Annotator Agreement For Word Sense

Disambiguation. In *Proc. of SIGLEX Workshop On Standardizing Lexical Resources*.

Palmer, M., Gildea, D., Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, *31(1):*71-106.

Paggio, P. (2006). Information structure and pauses in a corpus of spoken Danish. In *Proc. of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstration*, Trento, Italy.

Uria, L., Estarrona, A., Aldezabal, I., Aranzabe, M.J., de Ilarraza, A.D, Iruskieta, M. (2009). Evaluation of the Syntactic Annotation in EPEC, the Reference Corpus for the Processing of Basque. In *Proc. of 10th CICLing.*

Xia, F., Rambow, O., Bhatt R., Palmer, M., Sharma, D.M. (2009). Towards a Multi-Representational Treebank. In *Proc. of the 7th International Workshop on Treebanks and Linguistic Theories (TLT 2009),* Groningen, Netherlands.