A Hybrid Dependency Parser for Bangla

Arnab Dhar, Sanjay Chatterji, Sudeshna Sarkar, Anupam Basu
Department of Computer Sc. & Engineering, Indian Institute of Technology, Kharagpur, India
Email: {arnabdhar,schatt,sudeshna,anupam}@cse.iitkgp.ernet.in

ABSTRACT

In this paper we describe a two-stage dependency parser for Bangla. In the first stage, we build a model using a Bangla dependency Treebank and subsequently this model is used to build a data driven Bangla parser. In the second stage, constraint based parsing has been used to modify the output of the data driven parser. This second stage module implements the Bangla specific constraints with the help of demand frames of Bangla verbs. The features of the words used in both these stages include morphological features like gender, number, person, etc., parts-of-speech tags, chunk tags and named entity tags. The evaluation results show that this two stage parser performs better than one stage parsers.

1 Introduction

Parsing is a method of analyzing the grammatical structures (phrase structure or dependency structure) of a sentence. In this paper we refer parsing to indicate dependency parsing. Parsing has many applications in natural language processing such as, machine translation, anaphora resolution, question answering, etc. Data driven and grammar driven approaches have been used for parsing.

In a data driven parser of a language, a corpus is manually annotated with dependency relations of that language and used to train a model. This manually annotated corpus is referred to as dependency Treebank. In this paper we refer relation and Treebank to indicate dependency relation and dependency Treebank, respectively. In the Bangla Treebank released in Tool Contest of ICON 2009 and 2010 the dependency relations are assigned between chunks where in a chunk, words are related with intra-chunk relations. The data driven parser identifies the inter-chunk relations in the sentence using the model.

Grammar driven parser uses some language specific rules to identify the relations between chunks. These rules can be considered as constraints for dependents of a head words. If an input dependent of a head word satisfies the constraint then the corresponding relation is assigned between the dependent and the head word.

The availability of Treebank and rules on Bangla dependency grammar are limited. Therefore both the data driven and grammar driven dependency parsers often make mistakes in identifying the dependency relations. In this paper, we describe our experiment on correcting the output of a baseline data driven Bangla parser using Bangla specific constraints.

Proceedings of the 10th Workshop on Asian Language Resources, pages 55–64,
COLING 2012, Mumbai, December 2012.

2 Related work

Bharati et al. (1993) has described a constraint based Hindi parser by applying the Paninian framework. They have shown that the Paninian framework gives an account of the relation between the vibhakti and karaka roles in Hindi sentences. Bharati et al. (2002) has also used the computational Paninian framework for parsing Hindi sentences. The Hindi parser is developed based on translating Hindi grammatical constraints to integer programming constraints. This framework may be adapted for other free word order languages like Bangla. The transformations used by them operate on the grammar rules and not on the parse structures. Two stage constraint based approach described in Bharati et al. (2009) shows how intra-clausal and inter-clausal dependency relations are identified for each token in the sentence at two different stages.

Some study on demand frame or verb frame has been attempted on Hindi by Begum et al. (2008). De et al. (2009a) developed 500 Bangla demand frames including mixed verbs, main verbs and their causative forms.

The release of three Indian language Treebanks for Hindi, Bengali and Telugu in the Tool Contests of ICON 2009 and ICON 2010 have encouraged several researchers to work on Indian language parsing. The results of the nine participants of 2009 Tool Contest and six participants of 2010 Tool Contest are summarized in Husain (2009b) and Husain et al. (2010), respectively.

The participants of ICON 2009 Tool Contest have used both the data driven, grammar driven and hybrid approaches. For example, Nivre (2009) have suggested an optimal feature set for Hindi, Bengali and Telugu to be used in MaltParser.A grammar driven approach (constraint based) for Bengali language parsing has been suggested by De et al. (2009b). The approach consists of three steps, first simplify complex and compound sentential structures into simple sentences, then parse the simple structures by satisfying the Karaka demands of the Demand Groups (Verb Groups), and finally rejoin such parsed structures with appropriate links and Karaka labels. They have used 500 demand frames to achieve the highest performance in Bangla parsing. A hybrid approach has been suggested by Chatterji et al. (2009) where data driven parser used as a baseline system and postprocessed using four hard constraints namely (a) TAM based root identification (b) Genitive Marker Based Possessive Relation Identification (c) Resolving "pof" miss identification and (d) Post-position and suffix marker based rules.

The participants of ICON 2010 Tool Contest have also experimented with data driven, grammar driven and hybrid approaches. Kolachina et al. (2010) used data driven parser (MaltParser) where a linear-time algorithm for projective structures has been used as a parsing algorithm. They have also used LIBSVM learner as a learning algorithm and employed the propagation of some features in order to incorporate such features during the parsing process. They have achieved highest overall performance for Bangla in ICON 2010 Tool Contest. A hybrid approach has been proposed by Ghosh et al. (2010) where data driven parser (MaltParser) has been used as a baseline system and based on the errors of the data driven parser they have implemented a set of parsing rules based on vibhakti information.

3 Two stage parsing system

3.1 Motivation

Data driven parser learns the model from the Treebank of the language. Based on this model the parser identifies the dependency relations in the test sentences. If the size of Treebank of a language is limited, then the data driven parser of that language ought to make mistakes.

In Bangla sentences, the positions of the dependents of a head word are relatively less rigid. Both the dependents and the head words (including the root of the sentence) are often dropped in the Bangla sentences. Therefore, sometimes it is not possible for a data driven Bangla parser to correctly identify the relation.

On the other hand, a constraint based parser identifies relations based on a set of rules consisting of features of dependents and head words. For example, De et al. (2009) have proposed a set of Bangla rules by preparing case frames of Bangla verbs. Rule-set can be incrementally enhanced based on the relations required in a particular application such as machine translation.

So, we propose a hybrid parser where the data driven parser gives the baseline parsing in the first stage and the constraint based parser of the second stage may modify the mistakes of the first stage. Hopefully, this parser will give us better performance than the data driven and rule based parsers.

3.2 Overview

In this paper, we propose a two stage parsing system for Bangla. In the first stage, a data driven parser identifies the dependency relations between chunks in a sentence using the model created from the Bangla Treebank released in ICON 2009. We analyze the relations wrongly identified by the data driven parser and identify rules to correct them.

Researchers have tried several techniques and rules for correcting the mistakes of data driven parsers as discussed in Section 2. Researchers have also used case frames for identifying the dependency relations. Case frames show high performance in identifying the dependency relations in Bangla sentences. In this paper we experiment on the effects of case frames in the correction of mistakes of data driven parser.

3.3 Data driven module

The chunks in the Treebank used in the data driven parser has following attributes.

1. HEADWORD: Head word of chunk
2. HEADROOT: Root of head word of chunk
3. MORPH: Morphological features of head word: gender, number, person, animacy, case and vibhakti
4. POS: Part-of-speech tag of head word
5. CHUNK: Chunk tag
6. DEPREL: Dependency relation of that chunk with another chunk

We use the Covington's algorithm as implemented in MaltParser by Nivre (2006, 2007, 2009) for statistically annotating the dependency relations in Bangla sentences. In this algorithm, partially processed part is used for annotating the unprocessed part of a sentence. A stack is used for holding partially processed chunks (tokens) and a buffer is used for holding unprocessed tokens. The unprocessed tokens are annotated based on the features of the processed and unprocessed tokens. We have selected following set of features based on some experiments.

1. A set of POS features over stack and buffer of length 4.
2. A set of WORD features over stack and buffer of length 2.
3. A set of ROOT features over stack and buffer of length 2.
4. A set of CHUNK features over stack and buffer of length 1.
5. A set of POS features over dependents and head of length 1.
6. A set of combinations of the POS and WORD features of length 1.
7. A set of DEPREL features over dependents of length 1.
8. A set of WORD features over dependents and head of length 1.
9. A set of CHUNK features over dependents and head of length 1.
10. A set of MORPH features over stack and buffer of length 2.

The data driven Bangla parser built using these features achieves accuracy of 75.65% (Label Attachment Score).

3.4 Analyzing the mistakes of data driven module

The data driven parser makes mistakes in identifying some relations and attachments. One example sentence with mistakes of the data driven parser is discussed below. The Roman transliteration in Itrans and English translation are also included.

আমাকে দিল্লি যেতে হবে।
(AmAke dilli yete habe.)
[Me Delhi go have-to]
I have to go to Delhi.

In this example, আমাকে (AmAke)[me] is the Subject (karta) and দিল্লি (dilli) [Delhi] is the Spatial Locative (sthanadhikaran) of the verb যেতে হবে (yete habe) [have to go]. The data driven parser wrongly identifies আমাকে (AmAke) [me] as Object (karma) and দিল্লি (dilli) [Delhi] as Subject (karta). But, this non-transitive verb does not have Object (karma). Rather an animate noun may be used as Subject (karta). Similarly, this verb can't have 0-ending Subject (karta). Rather this noun may be used as Spatial Locative (sthanadhikaran).

This analysis shows that the mistakes made by the data driven parser in identifying the relations can be rectified using the grammar driven rules. We concentrate on correcting the mistakes of the data driven parser in identifying the following relations.

- karta [subject] (k1) and its subcategories
- karma [object] (k2) and its subcategories
- adhikarana [locative] (k7) and its subcategories
- part-of (pof)
- Verb Modifier (vmod)

3.5 Grammar Driven Module

A Grammar driven module is used as a postprocessor in the proposed hybrid system. This module corrects the mistakes made by the Bangla data driven parser by implementing some Bangla specific constraints. We represent these constraints in tabular way. These tabular representations are referred to as demand frames. The table for a verb contains possible relations with its dependents, the necessity of the relations and the features of the corresponding dependents. The necessity of a relation can be mandatory or desirable. The features we store in the demand frames are the vibhaktis, the part-of-speech tags, the named entity tags and the animacy.

For each Tense, Aspect, Modality (TAM) and certain other features of the Bangla verbs we build a set of transformation rules. All the verbs which have same feature follow same set of transformation rules. The transformation rules (with respect to a particular verb feature) identify the required changes in the basic demand frames of verb roots.

Basic demand frame for the Bangla verb root যাওয়া (yAoYA) [go] and transformation rule for the TAM তে_হবে (te_habe) [have-to] are represented in Table 1 and 2, respectively.

Dependency relation	Necessity	Vibhakti	Lexical type	NET	Semantic class
Subject (karta)	M	0	NN|NNP|PRP	0|PERSON	Animate|Inanimate
Spatial Locative (sthanadhikaran)	D	0|এ(e)|য়(Ya)|তে(te)	NN|NNP|PRP	0|LOCATION	0
Temporal Locative (kaladhikaran)	D	0|এ(e)|য়(Ya)|পর(para)	PRP|NN	0|TIMEX	0

TABLE 1 – Basic demand frame for verb যাওয়া (yAoYA) [go]
M: Mandatory, D: Desirable, NN:Noun, NNP:Proper noun, PRP:Pronoun

In Table 1 features of three dependents namely Subject (karta), Spatial Locative (sthanadhikaran), and Temporal Locative (kaladhikaran) for the verb root যাওয়া (yAoYA) [go] are shown. Subject (karta) is mandatory (M) and the other two dependents are desirable (D) for this verb. The possible values of the features are separated by | (pipe) symbol. Zero (0) indicates that the corresponding value of the feature is either Null or unknown.

Dependency relation	Necessity	Vibhakti	Lexical type	NET	Semantic class
Subject (karta)	M	কে(ke)	NN|NNP|PRP	0|PERSON	Animate|Inanimate

TABLE 2 – Transformation rule for the TAM তে_হবে (te_habe) [have-to].
M: Mandatory, NN: Noun, NNP: Proper noun, PRP: Pronoun

The features of Subject (karta) may change if the verb has TAM তে_হবে (te_habe) [have-to]. According to Table 2, the karta of this verb must have কে (ke) vibhakti.

Transformed demand frame for the verb token যেতে হবে (yete habe) [have to go] as shown in Table 3 is prepared from the basic demand frame of Table 1 and the transformation rule of Table 2. The vibhakti of Subject in the basic demand frame is transformed from 0 (Zero) to কে (ke) in the transformed demand frame.

The transformed demand frame is used to check the dependency relations of the verb token with its dependents as given by the data driven module. If a relation does not match with the corresponding entries in the transformed demand frame then we discard that relation. In this case, we find a relation for the dependent token of the discarded relation as follows.

Dependency relation	Necessity	Vibhakti	Lexical type	NET	Semantic class
Subject (karta)	M	কে(ke)	NN|NNP|PRP	0|PERSON	Animate|Inanimate
Spatial Locative (sthanadhikaran)	D	0|এ(e)|য়(Ya)|তে (te)	NN|NNP|PRP	0|LOCATION	0
Temporal Locative (kaladhikaran)	D	0|এ(e)|য়(Ya)|পর(para)	PRP|NN	0|TIMEX	0

TABLE 3 – Transformed demand frame for verb যেতে হবে (yete habe) [have to go]
M: Mandatory, D: Desirable, NN: Noun, NNP: Proper noun, PRP: Pronoun

The transformed demand frames for each verb (except the verb of the discarded relation) of the sentence are loaded. The features of the token are compared to these transformed demand frame entries. The nearest verb of the token whose transformed demand frame entries match with the features of the token is considered. The corresponding relation replaces the discarded relation.

3.6 Analyzing the effects of grammar driven module

The errors generated by the data driven parser for the example mentioned in Section 3.4 is corrected using the rules imposed by the following constraints.

- Based on the suffix (vibhakti) and semantic class value (animate or inanimate) of Subject (karta) of the verb token যেতে হবে (yete habe) [have to go] the relation of the word আমাকে (AmAke) [me] is changed from Object (karma) to Subject (karta).
- Based on the NET value of Spatial Lcative (sthanadhikaran) of the verb token যেতে হবে (yete habe) [have to go] the relation of দিল্লি (dilli) [Delhi] is changed from Subject (karta) to Spatial Locative (sthanadhikaran).

We show some more wrong outputs of the data driven parser and the effect of constraints on these outputs using dependency trees. The corrections in attachments and labels are shown using dotted lines and boldface, respectively.

1. বইটা তাকে দিয়ে বলল কাল পড়িস.
(ba;iTA tAke diYe balala kAla pa.Disa.)
[book him giving said tomorrow read.]
Giving him the book speaker says read tomorrow.

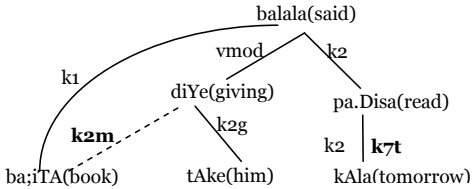
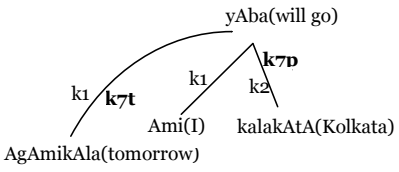


FIGURE 1 – Dependency tree for Example 1.

The corresponding dependency tree for Example 1 is shown in Figure 1. According to the transformed demand frame of the verb বলল (balala) [said] the Subject (k1) of this verb must be an animate noun. Therefore, the noun বইটা (ba;iTA) [book] can't be its Subject (k1). Then, according to the demand frame of the nearest verb দিয়ে (diYe) [giving] of this noun its Direct Object (k2m) is usually an inanimate noun. Again, according to the transformed demand frame of the verb পড়িস (pa.Disa) [read] a temporal noun can't be the karma of that verb. The temporal noun can be Temporal Locative (k7t) of that verb. Accordingly, we changed the attachments and relations given by the data driven parser.

2.1 আগামীকাল আমি কলকাতা যাব.
(AgAmikAla Ami kalakAtA yAba.)
[tomorrow I Kolkata will-go]
Tomorrow I will go to Kolkata.

2.2 গতকাল আমি কলকাতা দেখে এলাম.
(gatakAla Ami kalakAtA dekhe elAma.)
[yesterday I Kolkata see came]
Yesterday I visited Kolkata.



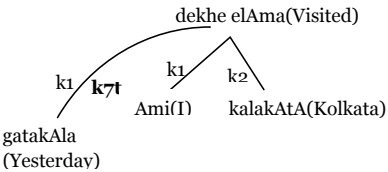


FIGURE 2(A) – Dependency tree for Example 2.1. FIGURE 2(B) – Dependency tree for Example 2.2.

The trees for Example 2.1 and Example 2.2 are shown in Figure 2(A) and 2(B), respectively. According to transformed demand frame of the verbs যাব (yAba) [will go] and দেখে_এলাম (dekhe elAma) [visited], a temporal noun can't be Subject (k1) of these verbs. This relation is changed to Temporal Locative (k7t). Again, according to the

transformed demand frame of the verb যাব (yAba) [will go], a spatial noun can't be Object (k2) and it is changed to Spatial Locative (k7p).

4 Evaluation Results

The data driven parser used in our task is trained on a Bangla Treebank containing 1200 sentences with an average length of 10.52 words. We tested the performance of this parser on a test data of 150 sentences.

In the grammar driven parser basic demand frames are prepared for 312 Bangla verbs and transformation rules for 12 Bangla verb features. Accuracy achieved by the data driven parser and the hybrid parser (data driven parser followed by grammar driven parser) are shown in Table 4. The table also contains the results achieved by other researchers for Bengali parsing on similar data-set. De et al. (2009) used a set of 500 Bangla demand frames and constraint based approach on the Bangla Treebank of ICON 2009. Kolachina et al. (2010) used MaltParser and various blended systems on the Bangla Treebank of ICON 2010.

	LAS	**UAS**	**LA**
Data Driven Parser	75.13	89.18	78.46
Hybrid Parser	80.35	89.63	84.20
De et al.	79.81	90.32	81.27
Kolachina et al.	75.65	88.14	78.67

TABLE 4 – Parser Evaluation Results.
LAS: Label Attachment Score, UAS: Unlabeled Attachment Score, LA: Label Accuracy.

The main improvements are achieved in the relations Subject (k1), Object (k2), Locative (k7) and Relation (r6). Table 5 shows the changes on the precision and recall of the label attachment scores of these relations.

	Subject		**Object**		**Locative**		**Relation**	
	R	P	R	P	R	P	R	P
Data Driven Parser	75.30	69.83	71.76	65.28	68.75	71.96	85.37	85.37
Hybrid Parser	86.06	83.04	82.44	78.83	77.27	73.91	89.02	83.91

TABLE 5 – Recall (R) and Precision (P) of subject, object, locative, and relation.

5 Conclusion

A two stage hybrid framework for dependency parsing of Bangla sentences is presented in this paper. In the first stage a data driven Bangla parser is developed using the experimentally calculated optimal features. We have developed a set of rules (called demand frames) for Bangla verbs. In the second stage, this demand frame based parser rectifies the mistakes in identifying the relations by the data driven parser.

More Bangla specific grammar rules may be developed for better performance of this framework. The performance of this framework can be tested for the parsing of sentences of other Indian language.

Acknowledgement

This work is partially supported by the ILMT project sponsored by TDIL program of MCIT, Govt. of India. We would like to thank all the members in Communication Empowerment Lab, IIT Kharagpur.

References

Begum, R., Husain, S., Bai, L., and Sharma, D. M. (2008). *Developing Verb Frames for Hindi*, In Proceedings LREC 2008.

Bharati, A., and Sangal, R. (1993). *Parsing Free Word Order Languages in the Paninian Framework*. In Proceedings of ACL:93.

Bharati, A., Sangal, R., and Reddy, T. P. (2002). *A Constraint Based Parser Using Integer Programming*. In Proceedings of ICON-2002.

Bharati, A., Husain, S., Vijay, M., Deepak, K., Sharma, D. M., and Sangal, R. (2009). *Constraint Based Hybrid Approach to Parsing Indian Languages*. In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23), Hong Kong.

Chatterji, S., Sonare, P., Sarkar, S., and Roy, D. (2009). *Grammar Driven Rules for Hybrid Bengali Dependency Parsing*. In Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India.

De, S., Dhar, A., and Garain, U. (2009a). *Karaka Frames and Their Transformations for Bangla Verbs*. In 31st All-India Conference of Linguists, Hyderabad, India.

De, S., Dhar, A., and Garain, U. (2009b). *Structure Simplification and Demand Satisfaction Approach toDependency Parsing in Bangla*. In Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India.

Gadde, P., Jindal, K., Husain, S., Sharma, D.M., Sangal, R. (2010). *Improving Data Driven Dependency Parsing using Clausal Information*. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 657–660, Los Angeles, California.

Ghosh, A., Das, A., Bhaskar, P. and Bandyopadhyay, S. (2010). Bengali Parsing System at ICON NLP Tool Contest 2010. In Proc of ICON-2010 tools contest on Indian language dependency parsing. Kharagpur, India.

Husain, S., Gadde, P., Ambati, B. R., Sharma, D., Sangal, R. (2009a). *A Modular Cascaded Approach to Complete Parsing*. IALP 2009, pages 141-146.

Husain, S. (2009b). *Dependency Parsers for Indian Languages*. In Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India.

Husain, S., Mannem, P., Ambati, B., and Gadde, P. (2010) *The ICON-2010 Tools Contest on Indian Language Dependency Parsing*. In Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing, Kharagpur, India.

Husain, S., Gadde, P., Nivre, J., Sangal, R. (2011). *Clausal parsing helps data-driven dependency parsing*: *Experiments with Hindi*. In Proceedings of IJCNLP 2011.

Kolachina, S., Kolachina, P., Agarwal, M., and Husain, S. (2010).*Experiments with MaltParser for parsingIndian Languages*. In Proc of *ICON-2010 tools contest on Indian language dependency parsing*. Kharagpur, India.

Mcdonald, R. (2007). *Characterizing the errors of data-driven dependency parsing models*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning.

Nivre, J., Hall, J., and Nilsson J. (2006). *MaltParser: A Data-Driven Parser-Generator for Dependency Parsing*. In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006), Genoa, Italy, pages 2216-2219.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov S., and Marsi, E. (2007). *MaltParser: A language-independent system for data-driven dependency parsing*. Natural Language Engineering, 13(2), pages 95-135.

Nivre, J. (2009). *Parsing Indian Languages with MaltParser*. In Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, Hyderabad, India.