# GuiTAR-based Pronominal Anaphora Resolution in Bengali

**Apurbalal Senapati**
Indian Statistical Institute
203, B.T.Road, Kolkata-700108, India
apurbalal.senapati@gmail.com

**Utpal Garain**
Indian Statistical Institute
203, B.T.Road, Kolkata-700108, India
utpal.garain@gmail.com

## Abstract

This paper attempts to use an off-the-shelf anaphora resolution (AR) system for Bengali. The language specific preprocessing modules of GuiTAR (v3.0.3) are identified and suitably designed for Bengali. Anaphora resolution module is also modified or replaced in order to realize different configurations of GuiTAR. Performance of each configuration is evaluated and experiment shows that the off-the-shelf AR system can be effectively used for Indic languages.

## 1 Introduction

Little computational linguistics research has been done for anaphora resolution (AR) in Indic languages. Notable research efforts in this area are conducted by Shobha et al. (2000), Prasad et al. (2000), Jain et al. (2004), Agrawal et al. (2007), Uppalapu et al. (2009). These works address AR problem in language like Hindi, some South Indian languages including Tamil. Dhar et al. (2008) reported a research on Bengali. Progress of the research through these works was difficult to quantify as most of the authors used their self-generated datasets and in some cases algorithms lack in required details to make them reproducible.

First rigorous effort was taken in ICON 2011 (ICON 2011) where a shared task was conducted on AR in three Indic languages (Hindi, Bengali, and Tamil). Training and test datasets were provided, both the machine learning (WEKA and SVM based classification) and rule-based approaches are used and the participating systems (4 teams for Bengali, and 2 teams each for Hindi and Tamil) were evaluated using five different metrics (MUC, $B^3$, CEAFM, CEAFE, and BLANC). However, no team attempted to reuse any of the off-the-shelf AR systems. This paper aims to explore this issue to investigate how far useful such a system is for AR in Indic languages. Bengali has been taken as the reference language and GuiTAR (Poesio, 2004) has been considered as the reference off-the-shelf system.

GuiTAR is primarily designed for English language and therefore, its direct application for Bengali is not possible for grammatical variations and resource limitations. Therefore, the central contribution of this paper is to develop required resources for Bengali and thereby providing them to GuiTAR for anaphora resolution. Our contribution also includes extension of the ICON2011 AR dataset for Bengali so that evaluation could be done on a bigger sized dataset. Finally, GuiTAR anaphora resolution module is replaced by a previously developed approach (which is primarily rule-based, Senapati, 2011; Senapati, 2012a) and performances of different configurations are compared.

## 2 Language specific issues in GuiTAR

GuiTAR has two major modules namely, preprocessing and anaphora resolution (Kabadjov, 2007). In both of these modules modifications are required to fit it to Bengali. Let's first identify the components in both of these two modules where replacement/modifications are needed.

**Pre-processing**: The purpose of this module is to make GuiTAR independent from input format specifications and variations. It takes as input in XML or text format. In case of text input, XML file generated by the LT-XML tool. The XML file contains the information like word boundaries (tokens), grammatical classes (part-of-speech), and chunking information. From the XML format MAS-XML (Minimum Anaphoric Syntax - XML) is produced to include minimal information namely, noun phrase boundaries, utterance boundaries, categories of pronoun, number information, gender information, etc. All these aspects are to be addressed for Bengali so that for a given input discourse in Bengali, MAS-XML file can be generated correctly. Next section explains how this issue.

**Anaphora resolution**: The GuiTAR system resolves four types of anaphoras. The *pronouns (personal and possessive)* are resolved by using



*Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 126–130,
Sofia, Bulgaria, August 4-9 2013. 

an implementation of MARS (Mitkov, 2002), whereas different algorithms are used for resolving *definite descriptions,* and *proper nouns*. In Mitkov's algorithm whenever a *pronoun* is to be resolved, it finds a list of potential antecedents within a given 'window' and checks three types of syntactic agreements (i.e., person, number and gender) between an antecedent and the pronoun. In case of more than one potential antecedent exists in the list it would be recursively filtered applying sequentially five different antecedent indicators (aggregate score, immediate reference, collocational pattern, indicating verbs and referential distance) until there is only one element in the list, i.e., the selected antecedent. We introduce suitable modifications in this module so that the same implementation of MARS can work for Bengali. This is explained in Sec. 4.

## 3 Bengali NLP Resources

Pronouns in Bengali has been studied before (linguistically by Majumdar, 2000; Sengupta, 2000 and for computational linguistics: Senapati, 2012a). Table-1 categorizes all pronouns (522 in number) available in Bengali as observed in a corpus (Bengali corpus, undated) of 35 million words.

| Category | Permissible Pronouns |
|---|---|
| Honorific Singular | তাঁর,তাঁকে, তিনি, তাঁরই, তিনিই,.. |
| Honorific Plural | তাঁরা, তাঁরাই, যাঁরা, উনারা,.. |
| 1st Person Singular | আমি, আমাকে, মোর,.. |
| 1st Person Plural | আমরা, আমাদেরকে, মোদের,.. |
| 2nd Person Singular | তোর, তোমার, আপনার,.. |
| 2nd Person Plural | তোরা, তোমরা, আপনারা,.. |
| 3rd Person Singular | এ, এর, ও, সে, তারও, তার,.. |
| 3rd Person Plural | এরা, ওরা, তারা, তাদের,.. |
| Reflexive Pronoun | নিজে, নিজেই, নিজেকে, নিজের,.. |

Table 1: Language resource

### 3.1 Number Acquisition for Nouns

In Bengali, a set of nominal suffixes (Bhattacharya, 1993) (inflections and classifier) are used to recognize the number (singular/plural) of noun. To identify the number of a noun, we check whether any of the nominal suffixes (indicating plurality) are attached with the noun. If found, the number of the noun is tagged as plural. From the corpus, we identified 17 such suffixes (e.g. দের /*der*, রা /*ra*, দিগের /*diger*, দিগকে /*digke*, গুলি /*guli*, etc.) which are used for number acquisition for nouns.

### 3.2 Honorificity of Nouns

The honorific agreement exists in Bengali. Honorificity of a noun is indicated by a word or expression with connotations conveying esteem or respect when used in addressing or referring to a person. In Bengali three degree of honorificity are observed for the second person and two for the third person (Majumdar, 2000; Sengupta, 2000). The second and third person pronouns have distinct forms for different degrees of honorificity. Honorificity information is applicable for proper nouns (person) and nouns indicating relations like father, mother, teacher, etc.

The honorificity information is identified by maintaining a list of terms which can be considered as honorific addressing terms (e.g. ভদ্রলোক/*bhadrolok*, বাবু/*babu*, ডঃ/*Dr.*, মহাশয়/*mohashoy*, ডা./*Dr.*, etc.). About 20 such terms are there in the list and we get these terms from analysis of the Bengali corpus. When these terms are used to add honorificity of a noun they appear either before or after the noun. Another additional way for identifying the honorificity information is to look at the inflection of the main verb which is inflected with ন/*n* (i.e. বলেন/bolen, করেন/koren etc.).

Honorificity is extracted during the preprocessing phase and added with the attribute *hon* = <*value*>. The value is set '*sup*' (superior i.e. highest degree of honor), '*neu*' (neutral i.e. medium degree of honor) or '*inf*' (inferior i.e. lowest degree of honor) based on their degree of honorificity. For pronouns, this information is available from the pronoun list (honorific singular and honorific plural) as shown in Table-1.

## 4 GuiTAR for Bengali

The following sections explain the modifications needed to configure GuiTAR for Bengali.

### 4.1 GuiTAR Preprocessing for Bengali

For getting part-of-speech information, the Stanford POS tagger has been retrained for Bengali language. The tagger is trained with about tagged 10,000 sentences and is found to produce about 92% accuracy while tested on 2,000 sentences. A rule based Bengali chunker (De, 2011) is used to get chunking information. NEIs and their classes (person, location, and organization) are tagged

manually (we did not get any Bengali NEI tool). After adding all these information, the input text is formatted into GuiTAR specified input XML file and is converted into MAS-XML. This file contains other syntactic information: person, types of pronouns, number and honorificity. Information on person and types of pronouns comes from Table-1. Number and honorificity are identified as explained before. Gender information has little role in Bengali anaphora resolution and hence is not considered. Types of pronouns are taken from Table-1.

### 4.2 GuiTAR-based Pronoun Resolution for Bengali

GuiTAR resolves pronouns using MARS approach (Mitkov, 2002) that makes use of several agreements (based on person, number and gender). Certain changes are required here as gender agreement has no role. This agreement has been replaced by the honorific agreement. Moreover, the way pronouns are divided in MARS implementation is not always relevant for Bengali pronouns. For example, we do not differentiate between personal and possessive pronouns but they are separately treated in MARS. In our case, we have only considered the personal and reflexive pronouns while applying MARS based implementation for anaphora resolution.

In case of more than one antecedent found, GuiTAR resolves it by using five antecedent indicators namely, aggregate score, immediate reference, collocational pattern, indicating verbs and referential distance. For Bengali, the indicating verb indicator has no role in filtering the antecedents and hence removed.

## 5 Data and data format

To evaluate the configured GuiTAR system the dataset provided by ICON 2011(ICON 2011) has been used. They provided annotated data (POS tagged, chunked and name entity tagged) for three Indian languages including Bengali. The annotated data is represented by a column format. Figure 1 shows a sample of the annotated data and the details description of the data is given in Table - 2.

```
story2.txt  0   0   সবশেষে NN  B-NP   O   -
story2.txt  0   1   তার  PRP B-NP   O   (13)
story2.txt  0   2   মনে  NN  B-NP   O   -
story2.txt  0   3   হলো  VM  B-VGF  O   -
story2.txt  0   4   এদিকে NN  B-NP   O   -
story2.txt  0   5   আর  QF  B-NP   O   -
```

Figure 2. ICON 2011 data format.

We have changed this format into GuiTAR specified XML format and finally checked/corrected manually. GuiTAR Preprocessor converts this XML into MAS-XML which looks like something as shown in Figure 3.

```
<s id="s81">
<ne gId="nv380" id="ne237">
<W Lpos="NN">সবশেষে</W></ne>
<ne gId="nv381" id="ne238" hon="neu"
AAcat="pers-pro" AAper="per3" AAnum="sing">
<nphead id="AAh54"><W Lpos="PRP">তার</W>
</nphead></ne>
<ne gId="nv382" id="ne239">
<W Lpos="NN">মনে</W></ne>
.....................
```

Figure 3. Sample GuiTAR MAS-XML file for Bengali text.

| Column | Type | Description |
|---|---|---|
| 1 | Document Id | Contains the file-name |
| 2 | Part number | File are divided into part numbered |
| 3 | Word number | Word index in the sentence |
| 4 | Word | Word itself |
| 4 | POS | POS of the word |
| 5 | Chunking | Chunking information using IOB format |
| 6 | NE tags | Name Entity Information is given |
| 7 | Description | Description |
| 8 | Co-reference | Co-reference information |

Table 2: Description of ICON 2011 data format

The ICON 2011 data contains nine texts from different domains (Tourism, Story, News article, Sports). We have extended this dataset by adding four more texts in the same format. Among these four pieces, three are short stories and one is taken from newspaper articles. Table 3 shows the distribution of pronouns in the whole test data set for Bengali.

| Data | ICON2011 | Extended |
|---|---|---|
| #text | 9 | 4 |
| #words | 22,531 | 4,923 |
| #pronouns | 1,325 | 322 |
| #anaphoric | 1,019 | 253 |

Table 3: Coverage of ICON 2011 dataset

## 6 Evaluation

The modified GuiTAR system has been evaluated by the dataset as described above. The dataset contains 1647 pronouns out of them 706 are personal pronouns (including reflexive pronouns). As the MARS in GuiTAR resolves only personal pronouns, we have used only these personal pronouns for evaluation. Three different systems are configured as described below:

System-1 (Baseline): A baseline system is configured by considering the most recent noun phrase as the referent of a pronoun (the first noun phrase in the backward direction is the antecedent of a pronoun).

System-2 (GuiTAR with MARS): In this configuration, GuiTAR is used with the modifications (as described in Sec. 4.1) in its preprocessing module and the modified MARS (as described in Sec. 4.2) is used for pronominal anaphora resolution (PAR).

System-3 (GuiTAR with new a PAR module): Under this configuration, GuiTAR is used with the modifications (as described in Sec. 4.1) in its pre-processing module but MARS is replaced by a previously developed system (Senapati, 2011; Senapati, 2012a) for pronominal anaphora resolution in Bengali. This is basically a rule-based system. For every noun phrase (i.e. a possible antecedent) the method first maintains a list of possible pronouns which the antecedent could attach with (note that any noun phrase cannot be referred by any pronoun). On encountering a pronoun, the method searches for the antecedents for which the pronoun is in the respective pronoun-lists. If there is more than one such antecedent, a set of rules is applied to resolve. The approach for applying the rules is similar to the one proposed by Baldwin (1997).

The evaluation has used five metrics namely, MUC, $B^3$, CEAFM, CEAFE and BLANC. The experimental results are reported in Table 4. Results show that GuiTAR with MARS gives better result than the situation where the most recent antecedent is picked (i.e. the baseline system). This improvement is statistically significant ($p<0.03$ in a two-sided t-test). When MARS is replaced by system-3, further improvement is achieved which is also statistically significant ($p<0.01$).

### 6.1 Error analysis

Analysis of errors shows that errors in number acquisition and identification of the honorificity are two major errors during preprocessing phase. These errors propagate and result in further errors during resolution. Resolution process itself introduces some new errors. For example, some Bengali personal pronouns are ambiguous (sometimes they are anaphoric whereas in other cases they may appear as non-anaphoric too). তার/tar, সে/se are two examples of such pronouns in Bengali (Senapati, 2012b) and the present resolution system is not able to resolve such cases.

| **System** | | **System-1 (Baseline)** | **GuiTAR** | |
|---|---|---|---|---|
| **Metric** | | | **System-2 (MARS)** | **System-3** |
| MUC | P | 0.453 | 0.516 | 0.538 |
| | R | 0.550 | 0.536 | 0.579 |
| | F1 | 0.497 | 0.526 | 0.558 |
| $B^3$ | P | 0.766 | 0.828 | 0.921 |
| | R | 0.771 | 0.824 | 0.911 |
| | F1 | 0.769 | 0.826 | 0.916 |
| CEAFM | P | 0.785 | 0.800 | 0.885 |
| | R | 0.632 | 0.622 | 0.784 |
| | F1 | 0.700 | 0.700 | 0.832 |
| CEAFE | P | 0.797 | 0.825 | 0.921 |
| | R | 0.552 | 0.571 | 0.731 |
| | F1 | 0.652 | 0.675 | 0.815 |
| BLANC | P | 0.688 | 0.700 | 0.732 |
| | R | 0.735 | 0.736 | 0.741 |
| | F1 | 0.711 | 0.718 | 0.736 |
| Avg. | F1 | 0.666 | 0.689 | 0.771 |

Table 4: Experimental results

## 7 Conclusion

The present experiment shows that GuiTAR which is one of the off-the-shelf anaphora resolution systems can be effectively configured for Bengali. Basic NLP information required by GuiTAR pre-processing module has been supplied mostly through automatic tools. A suitable tool is needed for NEI in Bengali. This can be explored in future. It is also revealed that MARS based implementation in GuiTAR is not very suitable for Bengali because the antecedent indicators used by MARS are probably not very effective for Bengali. Suitably designed rule based system could produce better result as shown in the experiment. Addition of other resolution algorithms is definitely a future extension of this study. Resolution of non-personal pronouns (which were not considered here) would be addressed next. In future, the similar experiment can be easily extended to other Indic languages (especially for Hindi and Tamil for which annotated data is available).

## References


Agarwal, S., Srivastava, M., Agarwal, P., Sanyal, R. 2007. Anaphora Resolution in Hindi Documents, in Proc. Natural Language Processing and Knowledge Engineering (IEEE NLP-KE), Beijing, China.

Baldwin, B. 1997. *CogNIAC: high precision coreference with limited knowledge and linguistic resources*, In ACL/EACL workshop on Operational factors in practical, robust anaphora resolution, pages 38- 45, Madrid, Spain.

Bengali Corpus. *TDIL Corpus in Unicode*, http://www.isical.ac.in/~lru/downloadCorpus.html.

Bhattacharya, T. and Dasgupta, P. 1993. *Classifiers, word order and definiteness in Bengali*, In Proceedings of the Seminar on Word Order. Osmania University, Hyderabad, India.

Dhar, A. and Garain, U. 2008. *A method for pronominal anaphora resolution in Bengali*, In Proc. of 6th Int. Conf. on Natural Language Processing (ICON), Student paper competition section, Pune, India.

De, S., Dhar, A., Biswas, S. and Garain, U. 2011. *On Development and Evaluation of a Chunker for Bangla*, In Proc. 2nd Int. Conf. on Emerging Applications of Information Technology (EAIT), pp. 321-324, Kolkata, India.

ICON. 2011. NLP Tools Contest: Anaphora Resolution in Indian Languages, In 9th Int. Conf. on Natural Language Processing (ICON), Chennai, India.

Jain, P., M R. Mital, S. Kumar, A. Mukerjee, and A. M. Raina. 2004. *Anaphora Resolution in Multi-Person Dialogues*, in Proc. 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, Massachusetts, USA.

Kabadjov, M.A. 2007. *A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification*, PhD thesis, Department of Computer Science, University of Essex.

Majumdar, A. 2000. *Studies in the Anaphoric Relations in Bengali*, Publisher: Subarnarekha, India.

Mitkov, R. 2002. *Anaphora Resolution*. Longman.

Poesio, M. and Kabadjov, M.A. 2004. *A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation*, in LREC 2004.

Prasad, R. and Strube, M. 2000. *Discourse salience and pronoun resolution in Hindi*, in Penn Working Papers in Linguistics, pp. 189-208.

Sengupta, G. 2000. *Lexical anaphors and pronouns in Bnagla*, Lexical Anaphors and Pronouns in Selected South Asian Langauges: A Principled Typology (Eds. B.C. Lust, K. Wali, J.W. Gair, and K.V. Subbarao), pp. 277-280, Publisher: Mouton de Gruyter, Berlin, New York.

Senapati, A. and Garain, U. 2011. *Anaphora Resolution System for Bengali by Pronoun Emitting Approach*, in Proc. NLP Tool Contest, 9th Int. Conf. on Natural Language Processing (ICON), Chennai, India.

Senapati, A. and Garain, U. 2012a. *Anaphora Resolution in Bengali using global discourse knowledge*, In Int. Conf. of Asian Language Processing (IALP), Hanoi, Vietnam.

Senapati, A. and Garain, U. 2012b. *Identification of Anaphoric **tAr*** (তার) *and **se*** (সে) *in Bengali*, In Proc. 34th All India Conference of Linguists (AICL), Shillong, India.

Sobha, L. and Patnaik, B.N.Patnaik. 2000. *Vasisth: An Anaphora Resolution System For Indian Languages*, In Proc. Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications (ACIDCA), Monastir, Tunisia.

Uppalapu, B. and Sharma, D.M. (2009). *Pronoun Resolution For Hindi*, in DAARC 2009.