A rule based system for Pronominal Anaphora Resolution in Bengali

Sayar Ghosh Roy

●●●

Abstract

In this project, we aim to build an anaphora resolution system for Bengali utilizing linguistic features for disambiguation of possible antecedents of a particular anaphor. The system accepts a sentence in UTF8 Bengali script as its input, identifies the pronominal anaphors and detects all possible antecedents for that particular anaphor. Using a set of hierarchical rules, it disambiguates and selects the best contextual antecedent from the set of possibilities. The results are illustrated on a manually tagged dataset and expected performance on real world data is discussed.

Related Work

Most of the pre-existing work performing anaphora resolution for Bengali relied on modifying an existing anaphora resolution framework to suit the likes of the Bengali language. Apurbalal Senapati and Utpal Garain built a system using the GuiTAR framework by fine-tuning the parameters for the Bengali language.[1] This work was conducted at the Indian Statistical Institute, Kolkata in 2013. Features such as Part-Of-Speech, Chunk and Named Entity information was used. They reported an accuracy of 77% on the ICON-2011 coreference resolution dataset. IIT Patna in collaboration

[1] Reference 1

with University of Trento, Italy, produced a paper[2] in 2013 in which they presented their very first attempt on anaphora resolution for a resource poor language. A state-of-the-art system, BART, which was originally developed for English was adapted for the purpose of Bengali. Overall performance of coreference resolution greatly depends on the high accurate mention detectors. They developed a number of models based on the heuristics used as well as on the particular machine learning paradigms employed. Thereafter, a series of experiments for adapting BART for Bengali was performed. The evaluation shows, a language-dependent system (designed primarily for English) can achieve a good performance level when re-trained and tested on a new language with proper subsets of features. The system produced the recall, precision and F-measure values of 56.00%, 46.50% and 50.80%, respectively. IIT Kharagpur conducted a study in this regard[3] and analyzed a data driven approach for Anaphora resolution of three Indian languages: Bengali, Hindi, and Tamil. The work consisted of two steps: identifying markables and links. Markable identification was done using Conditional Random Fields. The identifications of links between markables was done using Decision Tree Algorithm. They used the '*RandomTreeAlgorithm'* in Weka and reported an average f-value score of 66% for Bengali.

Rule Based Anaphora Systems

Singla et. al at Thapar University, Patiala built a rule based model for Anaphora resolution in Hindi exploiting dependency information.[4] Karaka labels on Hindi sentences were used to classify pronominal anaphors into classes including personal, reflexive, indefinite, relative and place pronouns. The dependency labels led to identification of antecedents using a set of tree traversal rules. Note that such a system could not be adapted for Bengali due to lack of robust dependency taggers. A dependency

[2] Reference 2
[3] Reference 3
[4] Reference 4

tagger for Bengali has been constructed at ISI[5] but it requires a specific format of input including 5 fields of information for each word in a sentence. It also suffers from improper use of WX notation. At present, LTRC[6], IIIT Hyderabad is building a dependency parser for Bengali which is not in a usable state at the very moment.

Key Idea behind my System

The International Journal of Computer Applications (IJCA), Foundation of Computer Science (FCS), NY, USA saw a study of Anaphora resolution in Bengali using rule based methods in November, 2016.[7] This work by Tazbeea Tazakka, Md. Asifuzzaman, Sabir Ismail at the Shahjalal University of Science and Technology, Bangladesh saw a system in which linguistic features such as Part-Of-Speech, number, gender, use of honorifics and person were used to resolve anaphora for pre-tagged Bengali text tailored specifically for the purpose of the experiment and achieved an accuracy of 80% on the same.

The System Implemented

The system accepts a sentence in Bengali UTF-8 script, identifies the pronominal anaphors and all possible antecedents for the same. It uses a set of rule based hierarchical constraints to prune out certain antecedents and reports the best possible antecedent in the given context, thereby resolving anaphora. We cannot use anything similar to the Hobb’s algorithm since Bengali is Indo-Aryan and in that, it is a free word order language. Hence, constituency parsing does not help capture any hierarchical structure. The features used are:

[5] Indian Statistical Institute

[6] Language Technologies Research Centre

[7] Reference 5

I. Part-Of-Speech information : The POS tag not only identifies pronouns and nouns, but also helps us recognize named entities in text.

II. Number : The pronoun and its antecedent must have number-agreement i.e references to singular and plural entities must tally with the antecedents themselves.

III. Person : Agreement of person helps classify one particular antecedent as more probable than another given other features are equivalent.

IV. Status : Honorifics used give us clues as to which particular person the referent is actually referring to.

V. Morphological : Morphological features are used to identify number in nouns and type of referent in case of pronouns.

My approach does not include the feature on gender used for other Indian Languages since Bengali does not capture gender information in pronouns. Instead, I incorporated morphological features into my system to help further disambiguate among the antecedents.

The following 3 divisions are observed among the set of pronouns in Bengali. Note that, only the head forms of the pronouns are illustrated here:

1. Number

singular_pronouns = ["তুমি", "তুই", "সে", "আপনি", "তিনি", "তার", "তোমার", "তোর", "আপনার", "আমার", "ওর"]

plural_pronouns = ["তোমরা", "তোরা", "তারা", "আপনারা", "তোমাদের", "তোদের", "আপনাদের", "আমাদের", "ওদের"]

Examples:

Singular : রাম বই পড়ছে । সে খুব ভালো । [Ram book read. He very good.]

Plural : অভিজ্ঞানরা দার্জিলিং বেড়াতে গেছে । তারা পরশু ফিরবে । [(Abhigyan &

others) Darjeeling tour gone. They (day after tomorrow) return.]

2. Person

first_person_pronouns = ["আমি", "আমার"]

second_person_pronouns = ["তুমি", "তুই", "আপনি", "তোমার", "তোর", "আপনার"]

third_person_pronouns = ["সে", "তিনি", "তার", "তারা", "ওদের", "ওর"]

Examples:

Third Person : শেলি ছবি আঁকছে । সেটি রঙিন । [Shelly picture drawing. It colorful.]

First Person : আমার নাম যদু। আমি কলকাতায় থাকি। [My name Jodu. Me (in Kolkata) stay.]

3. Status - These are the Honorific features

status_formal = ["আপনি", "তিনি", "আপনার", "আপনারা", "আপনাদের"]

status_informal = ["তুমি", "সে", "তার", "তোমার", "তোমরা", "তোমাদের"]

status_close = ["তুই", "তোর", "তোরা", "তোদের", "ওদের", "ওর"]

Examples:

Formal : রামবাবু সমাজের মাথা । ওনাকে সকলে ভয় পায় ।
[Ram-babu{honorific marker} society’s head. He everyone (feared by) feel.]

Informal : রাম বই পড়ছে । সে খুব ভালো । [Ram book read. He very good.]

Hence, the set of rules are applied in order to rule out the antecedents violating the constraints and finally selecting the referent.

Testing Sets

The correctness of the system has been verified using manually tagged Bengali data. Google’s transliteration tools for Indian Languages was used for creating this dataset. Some of the results and their implications will be discussed below. The system is then used to experiment on a large corpus of POS-tagged Bengali data collected from newspaper articles. This was provided to me by Ashutosh Ranjan who is a researcher at MT-NLP[8] laboratory, IIIT-H. The issues faced while using real world data will also be discussed.

Evaluating the System

The following cases of Success and Failure highlight the strengths and weaknesses of the system:

1

Input : রাম বই পড়ছে । সে খুব ভালো ।

Word by Word translation : Ram book read. He very good.

System Output:

Pronominal Anaphor : {'word': 'সে', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'রাম', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'বই', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Identified Antecedent : {'word': 'রাম', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

[8] Machine Translation - Natural Language Processing

Result : Success

Takeaway : Uses the morph features to identify that the reference is more likely to be to a named person.

2

Input : জুবের একটি বল নিয়ে খেলছে । সেটি গোলাকার ।

Word by Word translation : Zubair one ball with play. It round.

System Output:

Pronominal Anaphor : {'word': 'সেটি', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'জুবের', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'বল', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Identified Antecedent : {'word': 'বল', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Result : Success

Takeaway : Uses the morph features to identify that the reference is more likely to be to an inanimate entity and not a named person.

3

Input: অভিজ্ঞানরা দার্জিলিং বেড়াতে গেছে । তারা পরশু ফিরবে ।

Word by Word translation: (Abhigyan & others) Darjeeling tour gone. They (day after tomorrow) (will return).

System Output:

Pronominal Anaphor : {'word': 'তারা', 'POS': 'PRP', 'number': 'Plural', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'অভিজ্ঞানরা', 'POS': 'NP', 'number': 'Plural', 'person': 'Third', 'honor': 'Informal'}, {'word': 'দার্জিলিং', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Identified Antecedent : {'word': 'অভিজ্ঞানরা', 'POS': 'NP', 'number': 'Plural', 'person': 'Third', 'honor': 'Informal'}

Result: Success

Takeaway: Uses Number information to predict that the antecedent must be plural in number.

4

Input: শেলি ছবি আঁকছে । সেটি রঙিন ।

Word by Word translation: Shelly picture drawing. It colorful.

System Output:

Pronominal Anaphor : {'word': 'সেটি', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'শেলি', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'ছবি', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Identified Antecedent : {'word': 'ছবি', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Result: Success

Takeaway: Uses the morph features to identify that the reference is more likely to be to an inanimate entity and not a named person.

5

Input: রামবাবু সমাজের মাথা । ওনাকে সকলে ভয় পায় ।

Word by Word translation: Ram-babu(honorific marker) society’s head. He everyone (feared by) feel.

System Output:

Pronominal Anaphor : {'word': 'ওনাকে', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Formal'}

Set of Possible Antecedents : [{'word': 'রামবাবু', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Formal'}, {'word': 'সমাজের', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'মাথা', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Identified Antecedent : {'word': 'রামবাবু', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Formal'}

Result: Success

Takeaway: Uses the status and honorifics to identify that the individual being referred to will not be treated casually or informally.

6

Input: রাম আর যদু গান করছে । তারা ভালো গান গায় ।

Word by Word translation: Ram and Jadu song singing. They good song sing.

System Output:

Pronominal Anaphor : {'word': 'তারা', 'POS': 'PRP', 'number': 'Plural', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'রাম', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'যদু', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'গান', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

None of the Antecedents fit the Constraints

Result: Failure

Takeaway: The result should have been “রাম আর যদু”. The number constraint is being violated here. Hence, the system has to look beyond unit words and look at noun phrases which can combine multiple singular nouns into a plural entity. We do not have any constituency parser for Bengali which achieves this.

7

Input: বাঁদরটি নাচ করছিলো । তা দেখে ছেলেটি খুশি হলো ।

Word by Word translation: Monkey dance (was doing). That see boy happy became.

System Output:

Pronominal Anaphor : {'word': 'তা', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'বাঁদরটি', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

None of the Antecedents fit the Constraints

Result: Failure

Takeaway: The reference is to the entire act i.e the monkey dancing. An abstract anaphora resolver would capture the whole sentence as the antecedent - “বাঁদরটি নাচ করছিলো”. A simple pronominal resolver will not work here.

8

Input: রামের বাড়ি কলকাতায় । সেটা দেখতে বহু লোক যায় ।

Word by Word translation: Ram’s house (in Kolkata). That (to see) many people go.

System Output:

Pronominal Anaphor : {'word': 'সেটা', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'রামের', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'বাড়ি', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'কলকাতায়', 'POS': 'NP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Identified Antecedent : {'word': 'বাড়ি', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Result: Debatable

Takeaway: Correctly chooses the inanimate object i.e “house”. A more correct result would be “Ram’s house” i.e “রামের বাড়ি” instead. Again, one needs to capture multiword expressions in order to effectively do this.

9

Input: পাখির বাসা গাছে । সেটির রং নীল ।

Word by Word translation: Bird’s house (on tree). Its colour blue.

System Output:

Pronominal Anaphor : {'word': 'সেটির', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'পাখির', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'বাসা', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'গাছে', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Following Antecedents are Possible : [{'word': 'পাখির', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'বাসা', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'গাছে', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Result: Failure

Takeaway: All are equally possible. Need semantic information and world knowledge to go further and assign probabilities. The bird, the bird’s house or the tree - each of them possess the associated notion of color. They can all theoretically be blue. However, in order of likelihood, it is the bird that is blue, followed by the bird’s house followed by the tree.

10

Input: বাড়ির ওপর দিয়ে বিমান উড়ে গেল । সেটির গতি বিশাল ।

Word by Word translation: House’s top through plane fly. Its speed high.

System Output:

Pronominal Anaphor : {'word': 'সেটির', 'POS': 'PRP', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}

Set of Possible Antecedents : [{'word': 'বাড়ির', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'বিমান', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]

Following Antecedents are Possible : [{'word': 'বাড়ির', 'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}, {'word': 'বিমান',

```
'POS': 'NN', 'number': 'Singular', 'person': 'Third', 'honor': 'Informal'}]
```

Result: Failure

Takeaway: Needs semantic information and world knowledge. Needs to know that a typical house cannot fly. Linguistic information cannot achieve this.

Testing on given corpus

Statistics found:

# Sentences Processed : 9194

# Pronominal Anaphors Detected : 2059

# With possible Antecedents in Vicinity : 1309

# Without possible Antecedents in Vicinity : 750

We can clearly see that there are 750 cases of referents having no identifiable references in their vicinity. In a large document, the referring pronoun may refer to something being explained way before in the document or even way ahead, in case of cataphora. The present model suffers from localization and only works for small sets of sentences.

Also, a reference can be made to an entire paragraph. In literary works such as those of Ruskin Bond, a few paragraphs describe an event without explicitly naming it and if a pronoun is later used to refer to the event, the entire unit of the set of paragraphs should be captured as the antecedent. This is a far more difficult task and very little work has been done on this. This type of abstract anaphora resolution has been attempted using Bi-LSTMs and Siamese Nets in a mention ranking model for abstract anaphora resolution[9] which appears in the proceedings of EMNLP[10] 2017.

[9] Reference 6

[10] Empirical Methods in Natural Language Processing

Conclusion

Hence, for simple datasets, the system works well in identifying pronominal anaphors. It suffers from lack of semantic information, world knowledge and ability to capture hierarchical units of information which is required for solving the problem of resolving abstract anaphora. This study shows exactly how far a system with linguistic rules can go and illustrates the stepping off point after which the use of core knowledge becomes paramount. However, for a simple rule based system, it performs reasonably well often disambiguating a set of possible antecedents down to the most obvious and correct choice. It can be modified easily to capture hierarchical information by including those structures in the initial ambiguous set as well and thus, is really scalable. Capturing the semantic information with a global and domain specific knowledge base shall make the system extremely robust and reliable.

•••

References

[1] *https://www.aclweb.org/anthology/P13-2023*

[2] *http://www.scielo.org.mx/pdf//cys/v17n2/v17n2a4.pdf*

[3] *https://www.academia.edu/3275626/Anaphora_Resolution_for_Bengali_Hindi_and_Tamil_Using_RandomTree_Algorithm_in_Weka*

[4] *https://ieeexplore.ieee.org/document/8272666*

[5] *https://pdfs.semanticscholar.org/086c/b74ff9f2dcb915ec84e3be69fb567d859ff6.pdf*

[6] *https://arxiv.org/abs/1706.02256*