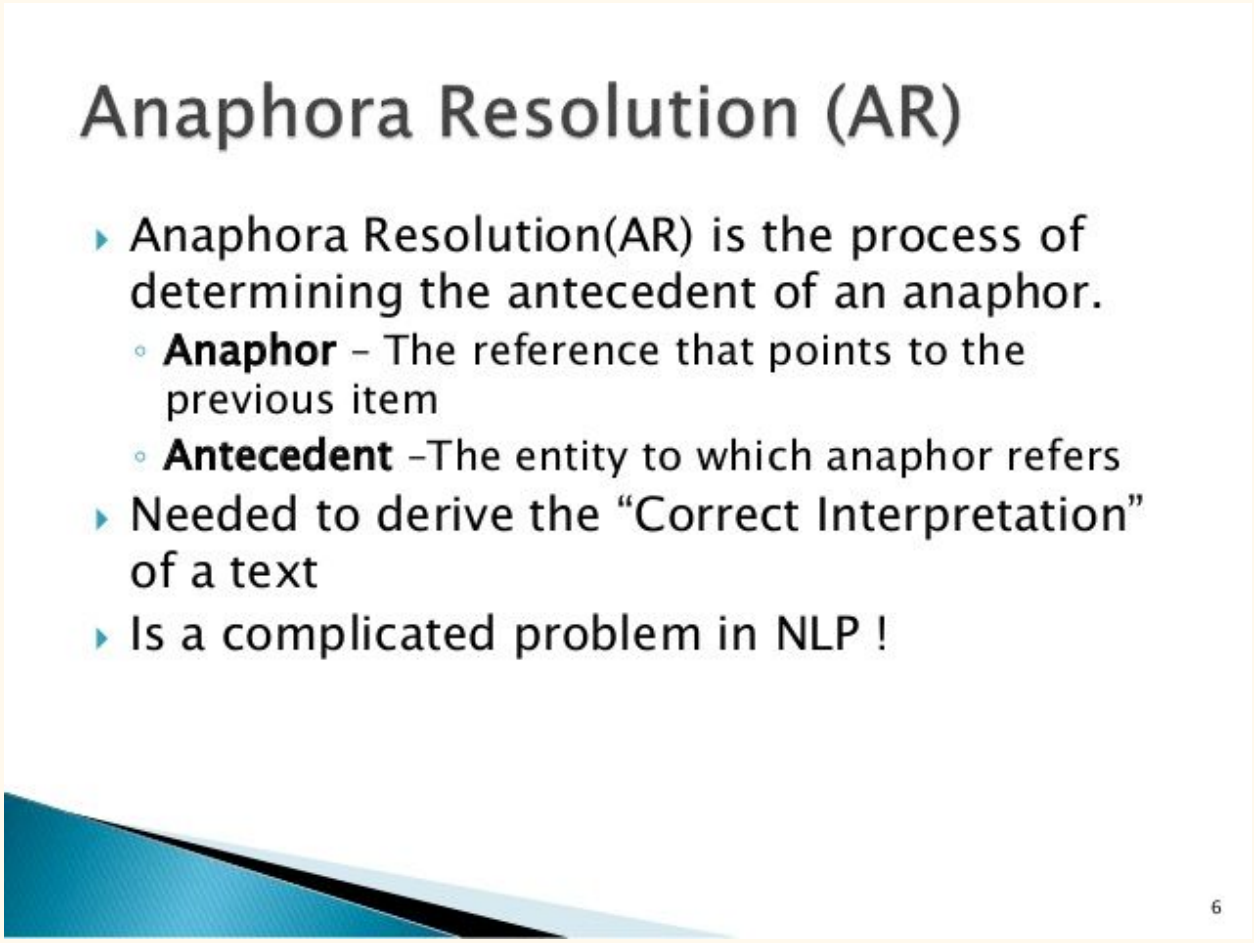
CL-II Project

A RULE-BASED APPROACH FOR ANAPHORA RESOLUTION IN BENGALI
FINAL REPORT

—

By Tathagata Raha



ABSTRACT

In this project, I am trying to build a rule-based anaphora resolution system for Bengali. We are focusing mainly on pronominal anaphora. The aim of the program will be to detect the pronominal anaphors and will select the most probable antecedent for a particular anaphora from a set of possible antecedents.

The program will input a piece of text which is a collection of few pre-tagged sentences in Bengali script and will output

- The pronominal anaphors in that text.
- Possible antecedents
- Antecedents corresponding to each anaphora

A comparative analysis has been done on a small tagged corpus highlighting its successes and failures.

RELATED WORK

Anaphora resolution has been tried in both rule-based and statistical approaches. A lot of research has been done in English starting from Hobbs search algorithm or centering theory to statistical approaches using neural nets.

In Bengali, most of the primitive anaphora resolution architectures depended on previously developed anaphora resolution frameworks. Apurbalal Senapati and Utpal Garain developed an anaphora resolution framework using the GuiTAR framework that was originally developed for English. They fine-tuned the parameters to make it suitable for Bengali and it resulted in an accuracy of 77p.c. on the ICON-2011 coreference resolution dataset. In 2013, a system known as BART which was originally developed for English was adapted for Bengali by a research team in IIT-Patna led by Asif Ekbal. But we cannot use these approaches cannot result in a greater accuracy because Indian languages have a free-word order which is different from other languages.

Rule-based systems

For Indian languages, a rule-based approach for Hindi has been developed by Praveen Dakwale, Vandan Mujadia and Dipti M Sharma of LTRC, IIIT Hyderabad. Here they used the dependency parsing information for disambiguating the anaphors and the references and they reported accuracy of 70p.c. However, due to the unavailability of a proper dependency parser for Bengali, I couldn't implement an approach based on dependency parsing.

Research paper referred for my approach

For my approach, I referred to a research paper "Anaphora Resolution in Bangla Language" by Tazbeea Tazakka, Md. Asifuzzaman and Sabir Ismail from Shahjalal University of Science and Technology which was published in the International Journal of Computer Applications (IJCA). This research paper uses a rule-based approach in which features like POS-tags, gender, number and honorifics to identify the pronominal anaphors and possible antecedents on pre-tagged data and achieved an accuracy of 77p.c. on the same.

IMPLEMENTATION

I have made a Python program that inputs a Bengali sentence in UTF-8 format and outputs

- The set of pronominal anaphoras
- The set of referents
- The set of possible referents for each anaphora

FEATURES OF REFERENTS AND PRONOUNS THAT WERE USED FOR DISAMBIGUATION OF ANAPHORS:

1. Parts-of-speech information- POStag not only helps in identifying the nouns and pronouns from the sentence but also in recognizing named entities.
2. Number- The pronouns should have number agreement i.e. singular pronouns should refer to singular pronouns and plural pronouns should refer to plural pronouns.
3. Person- Agreement of person helps classify one particular antecedent as more probable than another given other features are equivalent.
4. Honor- Honorifics actually help in identifying the person to whom it is being referred to.

5. Morphological features- They help in identifying what kind of referent they are referring to.

The implementation of my project is given in the steps below:

Take a properly tagged as input

Here at first, we take a Bengali sentence as input where each token will contain POS, number, person and honorific information.

For example:

রাম will be tagged as রাম/NP/S/T/I

- The first tag contains the word in UTF-8 format
- The second tag specifies the POStag of the word
- The third tag gives the number information.
  S-for singular
  P-for plural
  NA-for not applicable
- The fourth tag gives the person information
  F-for first person
  S-for second person
  T-for third person
  NA-for not applicable
- The fifth tag gives the honorific information
  F-Formal
  I-Informal
  C-Close

Parse the tagged sentence and process to make it usable in further steps

Then we break each token and we store the words in a dictionary along with their POStag, number, person and honour information.

```
def process_tokens(str):
    tokens=str.split(' ')
    processed_sentence=[]
    for i in tokens:
        word_dict={}
        temp_list=i.split('/')
        print(temp_list)
        if(len(temp_list)==1):
            word_dict['word'] = temp_list[0]
            word_dict['POS'] = "SYM"
            word_dict['number'] ="NA"
            word_dict['person'] ="NA"
            word_dict['honor'] ="NA"
        else:
            word_dict['word'] = temp_list[0]
            word_dict['POS'] = temp_list[1]
            word_dict['number'] =temp_list[2]
            word_dict['person'] = temp_list[3]
            word_dict['honor'] = temp_list[4]
        processed_sentence.append(word_dict)
    return(processed_sentence)
```

Identify the pronominal anaphors

Then we identify the pronominal anaphors i.e. wherever the POStag is PRP or PRP$

Identify the possible antecedents

Antecedents are nouns so we identify words with POStag NN or NP.

Classification of pronouns

Now all the pronouns are classified on the basis of

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"]

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"]
3. Honor
   Formal status: ["আপনি", "তিনি", "আপনার", "আপনারা", "আপনাদের"]
   Informal status: ["তুমি", "সে", "তার", "তোমার", "তোমরা", "তোমাদের"]
   Close status: ["তুই", "তোর", "তোরা", "তোদের", "ওদের", "ওর"]
4. Pronoun ending with "টা", "টি", "টির", "টিকে", "টাকে" refer to non-human entities.

Disambiguation

Now we find the antecedents that are fit for a particular anaphor by maintaining the following constraints:

1. The antecedent must have occurred before the anaphor.
2. The number, honor and person information should match with the anaphor.

ANALYSIS

Demonstration on the following sentences highlights the strengths and weaknesses of the system.

1.
   a. Input: রাম/NP/S/T/I বই/NN/S/T/I পড়ছে/VB/S/T/I । সে/PRP/S/T/I খব/JJ/S/T/I ভালো/JJ/S/T/I ।
   b. Word by word translation- Ram book read. He very good.
   c. System output:
      i. Pronominal anaphors: [{'word': 'সে', 'id': 5, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}]
      ii. Antecedents: [{'word': 'রাম', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'বই', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]
      iii. Identified antecedents: [{'word': 'রাম', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}]
   d. Verdict: Success

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

2.

a. Input: রামের/NP/S/T/I বল/NN/S/T/I আছে/VB/NA/NA/NA। সে/PRP/S/T/I ওটাকে/PRP/S/T/I খব/ADJ/NA/NA/NA ভালোবাসে/VB/NA/NA/NA। সেটি/PRP/S/T/I গোলাকার/VB/NA/NA/NA।

b. Word by word translation- Ram ball has. He that lot love. That circular.

c. System output:

i. Pronominal anaphors: [{'word': 'সে', 'id': 4, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'ওটাকে', 'id': 5, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'সেটি', 'id': 8, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}]

ii. Antecedents: [{'word': 'রামের', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'বল', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents for ‘সে’ : [{'word': 'রামের', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iv. Identified antecedents for ‘ওটাকে’ : [{'word': 'বল', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

v. Identified antecedents for ‘সেটি’ : [{'word': 'বল', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

d. Verdict: Success

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

3.

a. Input:রামবাবু/NP/S/T/F সমাজের/NN/S/T/I মাথা/NN/S/T/I । ওনাকে/PRP/S/T/F সকলে/QT/P/T/I ভয়/JJ/S/T/I পায়/VB/S/T/I ।

b. Word by word translation- Ram society head. Him everyone fears.

c. System output:

i. Pronominal anaphors: [{'word': 'ওনাকে', 'id': 5, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'F'}]

ii. Antecedents: [{'word': 'রামবাবু', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'F'}, {'word': 'সমাজের', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'মাথা', 'id': 3, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents: [{'word': 'রামবাবু', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'F'}]

d. Verdict:Success

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

4.

a. Input: মুকুন্দরা/NP/P/T/I দার্জিলিং/NP/S/T/I বেড়াতে/VB/S/T/I গেছে/VB/S/T/I । তারা/PRP/P/T/I পরশু/NN/S/T/I ফিরবে/VB/P/T/I ।

b. Word by word translation- Mukund(plural) Darjeeling went. They day after tomorrow come.

c. System output:

i. Pronominal anaphors: [{'word': 'তারা', 'id': 6, 'POS': 'PRP', 'number': 'P', 'person': 'T', 'honor': 'I'}]

ii. Antecedents: [{'word': 'মুকুন্দরা', 'id': 1, 'POS': 'NP', 'number': 'P', 'person': 'T', 'honor': 'I'}, {'word': 'দার্জিলিং', 'id': 2, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'পরশু', 'id': 7, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents: [{'word': 'মুকুন্দরা', 'id': 1, 'POS': 'NP', 'number': 'P', 'person': 'T', 'honor': 'I'}]

d. Verdict: Success

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

5.

a. Input: নিখিল/NP/S/T/I ছবি/NN/S/T/I আঁকছে/VB/S/T/I । সেটি/PRP/S/T/I রঙিন/JJ/S/T/I ।

b. Word by word translation-Nikhil art drawing. That colourful.

c. System output:

i. Pronominal anaphors: [{'word': 'সেটি', 'id': 5, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}]

ii. Antecedents: [{'word': 'নিখিল', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'ছবি', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents: [{'word': 'ছবি', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

d. Verdict: Success

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

6.

a. Input: পাখির/NN/S/T/I বাসা/NN/S/T/I গাছে/NN/S/T/I । সেটির/PRP/S/T/I রং/NN/S/T/I নীল/NN/S/T/I ।
b. Word by word translation- Bird nest tree. That colour blue
c. System output:
    i. Pronominal anaphors:[{'word': 'সেটির', 'id': 5, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}]
    ii. Antecedents: [{'word': 'পাখির', 'id': 1, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'বাসা', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'গাছে', 'id': 3, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'রং', 'id': 6, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'নীল', 'id': 7, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]
    iii. Identified antecedents: [{'word': 'পাখির', 'id': 1, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'বাসা', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'গাছে', 'id': 3, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]
d. Verdict: Debatable
e. Takeaway: All are equally possible. Need semantic information and world knowledge to go further and assign probabilities. The bird, the nest 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 nest followed by the tree.

7.

a. Input: রাম/NP/S/T/I আর/CONJ/S/T/I যদ/NP/S/T/I গান/NN/S/T/I করছে/VP/S/T/I । তারা/PRP/P/T/I ভালো/JJ/S/T/I গান/NN/S/T/I গায়/VB/S/T/I ।
b. Word by word translation- Ram and Jadu are singing. They good sing.
c. System output:
    i. Pronominal anaphors: [{'word': 'তারা', 'id': 7, 'POS': 'PRP', 'number': 'P', 'person': 'T', 'honor': 'I'}]
    ii. Antecedents: [{'word': 'রাম', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'যদ', 'id': 3, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'গান', 'id': 4, 'POS': 'NN', 'number': 'S', 'person': 'T',

'honor': 'I'}, {'word': 'গান', 'id': 9, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents: None

d. Verdict: Failure

e. Takeaway: The result should have been “রাম আর যদ” . 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.

8.

a. Input: রামের/NP/S/T/I বাড়ি/NN/S/T/I কলকাতায়/NP/S/T/I । সেটা/PRP/S/T/I দেখতে/VB/S/T/I বহু/JJ/P/T/I লোক/NN/S/T/I যায়/VB/S/T/I ।

b. Word by word translation-Ram’s house Kolkatay. That see many people

c. System output:

i. Pronominal anaphors: [{'word': 'সেটা', 'id': 5, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}]

ii. Antecedents: [{'word': 'রামের', 'id': 1, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'বাড়ি', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'কলকাতায়', 'id': 3, 'POS': 'NP', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'লোক', 'id': 8, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents: [{'word': 'বাড়ি', 'id': 2, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

d. Verdict: Failure

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

9.

a. Input: বাড়ির/NN/S/T/I ওপর/POST/S/T/I দিয়ে/VAUX/S/T/I বিমান/NN/S/T/I উড়ে/VB/S/T/I গেল/VAUX/S/T/I । সেটির/PRP/S/T/I গতি/NN/S/T/I বিশাল/JJ/S/T/I ।

b. Word by word translation- House above aeroplane goes. That speed very much.

c. System output:

i. Pronominal anaphors: [{'word': 'সেটির', 'id': 8, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}]

ii. Antecedents: [{'word': 'বাড়ির', 'id': 1, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'বিমান', 'id': 4, 'POS': 'NN', 'number': 'S', 'person':

'T', 'honor': 'I'}, {'word': 'গতি', 'id': 9, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents: [{'word': 'বাড়ির', 'id': 1, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'বিমান', 'id': 4, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

d. Verdict: Failure

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

10.

a. Input: বাঁদরটি/NN/S/T/I নাচ/VB/S/T/I করছিলো/VB/S/T/I । তা/PRP/S/T/I দেখে/VB/S/T/I ছেলেটি/NN/S/T/I খুশি/JJ/S/T/I হলো/VAUX/S/T/I ।

b. Word by word translation-Monkey dancing. That seeing boy happy.

c. System output:

i. Pronominal anaphors: [{'word': 'তা', 'id': 5, 'POS': 'PRP', 'number': 'S', 'person': 'T', 'honor': 'I'}]

ii. Antecedents: [{'word': 'বাঁদরটি', 'id': 1, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}, {'word': 'ছেলেটি', 'id': 7, 'POS': 'NN', 'number': 'S', 'person': 'T', 'honor': 'I'}]

iii. Identified antecedents: None

d. Verdict: Failure

e. 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 - “বাঁদরDž নাচ করিছেলা”. A simple pronominal resolver will not work here.

CHALLENGES

1. As gender information is not present in the pronouns, it will be difficult to distinguish between male and female antecedents.
2. Semantic information couldn’t be expressed always through the words (as in example 7)
3. This approach can only recognise data within a vicinity. It does not assign any priority to any word.
4. Sometimes, an anaphor can relate to a whole event.
5. Lack of real-world data as in example 6 and 9
6. Some pronominal anaphors can direct to verbs as in example 10

FUTURE WORK

1. Developing a dependency parser for Bengali and by using the depending information would enhance the performance to a great extent
2. Using neural network models like LSTMs and Siamese nets can help in giving a score to the possible antecedents and mark its relevance with an anaphor.
3. By capturing the semantic information for a global or domain-specific knowledge