Anaphora Resolution in Bangla Language

Tazbeea Tazakka
Department of CSE
Shahjalal University of Science and Technology

Md. Asifuzzaman
Department of CSE
Shahjalal University of Science and Technology

Sabir Ismail
Department of CSE
Shahjalal University of Science and Technology

ABSTRACT
In this paper we have discussed about our work of Anaphora resolution in Bangla. For solving this problem pronouns and verbs are considered as anaphor and their antecedent nouns. After trying several methods, first attempt was to find out if there is any relation among the classifications of pronouns and verb with those of nouns. That indicates to find out which types of pronouns refer to which type of nouns. But that process does not work as expected. This shows that, anaphor and antecedent in Bangla matches with each other in some factors that are also considered in other languages. Those are Number, Gender and Person. In case of personal pronouns in Bangla status of person can be differentiated such as Honorable, Normal and Negotiable. Experimenting with these factors proved that this method works well and give accuracy of around 80%. Though there are errors as, because of some nouns can have the same factors like the anaphor, they are not the antecedent in real.

General Terms
Natural language processing, Anaphora resolution, Reference resolution, Bangla Language Processing.

Keywords
Anaphora, Antecedent, Reference Resolution, Bangla Language Processing.

1. INTRODUCTION
Anaphora Resolution indicates a part of reference resolution. Reference resolution means which words or phrases refer to which previous or next words or phrases and how they are related in a sentence or in a paragraph or in a whole document. Anaphora stands for the word or phrase that is referring to other previous words or phrases.

হৃদয় তার বই হারিয়ে ফেলেছে।

Here "হৃদয়" is the antecedent and "তার" is the anaphor. The word "তার" refers back to "হৃদয়". Anaphor always refers backward. The referring word or anaphor does not have to be a pronoun. It can be pronoun, noun, noun phrase, verb phrase and adjectives. The antecedent can be noun, phrase or an event. Such as:

রিমা খেলছে।

একটি ছেলে ক্লাসে এসে ঢুকল। **তাকে** দেখে মনে হচ্ছে কিছু একটা খুঁজছে।

ছেলেটি বাড়ি ছেড়ে চলে গেল। সেটা তার বাবা-মায়ের জন্যে অনেক কষ্টের ছিল।

Anaphora Resolution is important for some reasons in Natural Language Processing. According to [1], it indicates the construction of discourse. At the sentence level it binds different elements of sentences together. It helps to understand and process language. As finding out reference is very difficult it is a challenge to NLP.

Reference [2] shows that by finding parts of a text containing a given topic it is possible to perform text summarization and information extraction more reliably. An implementation of anaphora resolution is found in question answering systems. So it can be easily understood that this is very important for Bangla Language also.

In the next chapter, related works of reference resolution is discussed. Then there is proposed method. In chapter 4, experimental data of the proposed methods is discussed and then result analysis is discussed. Finally the concluding parts containing the future scopes.

2. RELATED WORKS
As reference resolution is a very wide area. It can be done in different areas. Such as: Proper Noun Co reference: It is the easiest to find the correct answer and it is also very important for its many applications. Pronoun Co reference: It is the most thoroughly studied area in linguistics. Common Noun Coreference: It is the most complex area in linguistics.

Reference [2] shows that they have worked with pronoun Co reference. They used the Hobbs-Search Algorithm. In this method when a pronoun is found, then it backtracks to find its antecedent. First the previous sentence is searched, and then it goes on until the first of the text if the antecedent is not found.

As in [3] Information Structure of a sentence is used to solve the reference resolution problem when the sentence is not structured. Such as: I am going to show you – close the door please- how sentence is structured.

This problem is solved according to what the speaker is treating as information that is recoverable to the hearer and what he is treating as non-recoverable to the hearer.

As in [4] binding theory is applied at the earliest possible stage in processing and constrains all subsequent stages of reference resolution. This can be said from the binding-as-initial filter hypothesis (Nicol & Swinney, 1989).

For example:

Rahim thinks that **Karim** did the work **himself**.

According to the binding-as-initial filter Karim is chosen as the antecedent of himself, which is of course correct though the focus of the sentence was Rahim. Binding theory is really fruitful in the case of working with reciprocal or reflexive pronouns.

According to [5], pronouns and definite noun phrases are the most common for referring expressions. A new feature based on minimum edit distance between anaphor and antecedent was adopted which gave a significantly improved result in case of definite noun phrases and proper noun.

This method also includes features which differentiate the gender of a noun or pronoun. They also capture the difference among human, concrete objects and abstract objects.

Reference [1] tells us that they have developed an anaphora resolution system for Bangla using mention detection in the first stage using conditional random field (CRF) and in the second stage using BART, a system successfully used in different languages. They used some features for mention detection such as- context word, suffix and prefix of word, pronoun list. Then in the second stage they applied the BART method. But it was clear that this method did not work exactly as the other languages because Bangla has completely different characteristics from those languages. Nevertheless they used String matching, distance between antecedent and anaphora, co-reference chain as features of this method. It gave better results than the previous methods they have planned to progress with.

According to [6], they have developed a two stage pronoun reference system for Bangla, Hindi and Tamil. In the first stage they try to find the boundary of a markable item such as a single noun chunk or a group of noun, conjunction or adjectival chunk. Then in the second stage they have used a decision tree algorithm to identify antecedent-anaphoric relation.

According to [7], they have solved anaphora resolution for Bangla using a different process named GuiTAR which has two different modules preprocessing and anaphora resolution. The preprocessing stage makes GuiTAR independent from input format specifications and variations. This stage includes POS tagging, categorization of pronoun, etc. from the text given in XML. The second stage, pronoun resolution for Bangla, checks person, number and gender of noun. If more than one noun is found, then it applies more filters such as aggregate score, immediate reference and collocation pattern. This system has worked for them, but it shows so many errors in both the stages.

According to [8], anaphora resolution can be divided into two parts and they are 1.co-reference resolution, which specifies finding co-reference of full Noun Phrase, 2.anaphora resolution which indicates finding reference of a pronoun or reflexive. In this paper anaphora resolution is said to be in two broad categories and they are knowledge-rich approaches and knowledge-poor approaches. Knowledge-rich approaches are based on commonly observed heuristics about anaphoric phenomenon. These approaches need a full and correct input. So evaluation is carried out by hand on a small set of examples. The knowledge-poor approaches mainly concentrate on automated system which includes machine learning.

3. PROPOSED METHOD

As said in [8] working with pronoun resolution is chosen and as for the method knowledge-rich approach is preferred which needs to work manually and needs correct data. It is because if knowledge-poor system is used from the start, any error at any stage, for example at POS tagging, can lead us to erroneous result which will be difficult to solve.

According to [9] anaphora and antecedent matches on some different categories. In the following approach it is tried to find out if anaphor and antecedent match in some categories in Bangla also.

3.1 Number (বচন) as a factor

In English Language, number of the subject of a sentence or the number of the noun will have the same number in pronouns. In case of 3rd personal pronouns and 1st personal pronouns Singular and Plural Number can be distinguished. But it can't be distinguished in 2nd personal pronouns as shown in Table 1.

Table 1. List of pronouns according to number

Number	Pronoun
Singular	He, She, His, Her, Him, You, Your, I, Me, My, Himself, Herself, Myself
Plural	They, Their, Them, You, Your, We, Our, Themselves, Ourselves

In Bangla numbers (একবচন, বহুবচন) in 1st, 2nd and 3rd person can be distinguished as in Table 2.

Table 2. List of সর্বনাম according to একবচন and বহুবচন

বচন	সর্বনাম
একবচন	তুমি, তুই, সে, আপনি, তিনি, তাঁর, তোমার, তোর, আপনার, আমার, আমি
বহুবচন	তোমরা, তোরা, তারা, আপনারা, তাঁরা, তোমাদের, তোদের, আপনাদের, আমাদের

So this indicates that Number can be used as a factor in Bangla Language.

3.2 Gender (লিঙ্গ) as a factor

Gender is a very efficient factor in Anaphora Resolution of English Language. But in Bangla Language it is not that helpful. Because in English they use different pronouns for different gender such as "He" and "She". But Bangla only has "সে". So any differences according to gender can not be pointed in case of pronouns. But considering that it can be helpful in future, records of Gender were kept to find reference of Noun Phrases.

In this method classifications of Nouns and Pronouns are not used. Gender can differentiate among some of them. Gender is classified as Male, Female, Common, Undetermined. Here common indicates objects. That means the Gender "C" which actually indicates to বস্তুবাচক বিশেষ্য. So, it helps to find out antecedent of an anaphor.

3.3 Person (পুরুষ) as a factor

Personal Pronouns can also make a huge difference in case of anaphora resolution. So this can be also a factor.

3.4 Status of a person as a factor

While using 2nd and 3rd Person in Bangla there is another factor that could be helpful in many cases. Those are সাধারণার্থে and সম্মানার্থে .Using these factors increased the accuracy of the anaphora resolution. Some of those are listed in below.

Table 3. List of সর্বনাম according to সাধারণার্থে and সম্মানার্থে

পুরুষ	সর্বনাম
সাধারণার্থে	তুমি, তুই, তোমরা, তোরা, তোমার, তোমাদের, তোদের, সে, তারা, তাদের ইত্যাদি৷
সম্মানার্থে	আপনি, আপনারা, আপনাদের, তিনি, তাঁর, তাঁদের ইত্যাদি৷

Using these factors, along with parts of speech (বিশেষ্য, সর্বনাম) the most accurate answers were found so far. Most of the verbs have almost all the categories in them to be an anaphor according to this method. So verbs are also included in this method as anaphor also.

The best results has been got through this method.. But in this case data set was not same.

4. EXPERIMENT

4.1 Method one

Bangla language has a large number of pronouns. At first Bangla Pronouns were listed according to their classifications. So first of all steps were taken to proceed with a method when a pronoun is found in a sentence it looks for the previous noun.

4.2 Method two

The results of the first step showed that all the pronouns that are ব্যাক্তিবাচক সর্বনাম only refers back to জাতিবাচক বিশেষ্য / সমষ্টিবাচক বিশেষ্য .So next this process was included with the previous one.

4.3 Method three

As the previous method does not work with any other type of pronoun than ব্যাক্তিবাচক সর্বনাম, patterns like this were looked for among those types. But there was no such pattern that can be pointed out from the existing corpus. So the procedure of "Method One" was tried once again for the other types of pronoun.

4.4 Method four

Though ব্যাক্তিবাচক সর্বনাম refers to both জাতিবাচক বিশেষ্য and সমষ্টিবাচক বিশেষ্য the percentage of the second one is much less than the first one. So only referring back to জাতিবাচক বিশেষ্য was tried next.

4.5 Method five

For Method five another corpus was used. In this method five factors have been used to find out matches between nouns (Antecedent) and pronouns (Anaphor) or verbs (Anaphor). The experimental study of these factors are given below.

4.5.1 Number

"S" is used for একবচন and "P" for বহুবচন. But with these two types, verbs can't be marked out for most of the cases.

For example consider the verb "করেছে" and "করেছ"**.**

সে করেছে। তারা করেছে।

তুমি করেছ। তোমরা করেছ।

So another type of Number is included that is called "U" (Undetermined) which is used in those cases where anaphor can be used as both of the numbers as the examples given above.

4.5.2 Person

In case of person "FP", "SP" and "TP" were used for First Person, Second Person and Third Person consecutively which are just like Bangla. But in case of verbs as "অসমাপিকা ক্রিয়া" person of a verb cannot be determined.

For example:

আমি বইটি নিয়ে গেলাম।

তুমি বইটি নিয়ে যাও ।

সে বইটি নিয়ে গেল।

In this case the verb "নিয়ে" can be used in all the persons used in Bangla. So for this kind of problem "UP" is included as Undetermined Person in the tagging.

4.5.3 Status of a person

By finding out the person of a verb and pronoun does not determine all the basic properties of a pronoun or verb. In Bangla Language within Person there is another property that is the status of the person. Those are "সম্মানার্থে", "সাধারণার্থে" and "তুচ্ছার্থে".

আমি কাজটি করেছি।

আপনি কাজটি করেছেন।

তুমি কাজটি করেছ।

তুই কাজটি করেছিস।

সে কাজটি করেছে।

তিনি কাজটি করেছেন।

The verbs and pronouns is shown in a tabular form below.

Table 4. Distribution of pronoun and verb according to

পুরুষ	সম্মানার্থে	সাধারণার্থে	তুচ্ছার্থে
উত্তম (1^{st})		আমি ; করেছি	
মধ্যম (2^{nd})	আপনি; করেছেন	তুমি ; করেছ	তুই ; করেছিস
তৃতীয় (3^{rd})	তিনি ; করেছেন	তুমি ; করেছ	তুই ; করেছিস

From the above table it can be seen that "তুচ্ছার্থে" is used only in case of Second Person. But no noun phrase can be in this status. Noun can be only in "সম্মানার্থে" and "সাধারণার্থে" status. So, "তুচ্ছার্থে" is considered as "সাধারণার্থে".

In case of "অসমাপিকা ক্রিয়া" where the status of a person cannot be determined "U" is used for Undetermined. For "সাধারণার্থে" and "সম্মানার্থে", "N" and "H" are used consecutively. For example:

তুমি বইটি নিয়ে যাও।

আপনি বইটি নিয়ে যান।

তুই বইটি নিয়ে যা।

4.5.4 Gender

In case of Gender four tags have been used. They are "M", "F", "C" and "U" used consecutively for "পুংলিঙ্গ", "স্ত্রীলিঙ্গ", "ক্লীবলিঙ্গ" and "Undetermined".

In Bangla male and female cannot be differentiated in case of pronoun or verbal phrases. Such as:

রিমা তার চাবি হারিয়ে ফেলেছে।

হৃদয় তার চাবি হারিয়ে ফেলেছে।

In both cases "তার" and "হারিয়ে ফেলেছে" cannot be determined if it refers to a male or a female. So they will be considered as "U". But in case of noun that means when the anaphor is also a noun it may be determined sometimes. For example:

রিমা খুব ভাল মেয়ে।

In this case usually "রিমা" is a name of a girl. That means the gender can be determined here as "F". Afterwards "মেয়ে" also indicates to "F". From this it can be said that the word "মেয়ে" is referring back to "রিমা".

In case of "ক্লীবলিঙ্গ" "C" is used. This indicates materials or nouns other than person. For example:

জাহাজটি ঝড়ের মুখে পড়েছিল। কিন্তু সেটির কোন ক্ষতি হয়নি।

Here the pronoun "সেটির" does not stand for any person. It stands for a material, an object or an event. That means it stands for gender type "C" which same as the gender type of the noun "জাহাজ" in the previous sentence. So the pronoun refers back to this noun which is correct.

4.5.5 Parts of speech (Noun/Pronoun/Verb)

In this method only Noun, Pronoun and Verb are used among Parts of speech. The tags are like this: "NO" for Noun, "PN" for Pronoun and "V" for Verb.

The previous five factors are used to match with a pronoun or verb and its previous nouns. When the factors of the antecedent does not match any factor of the anaphor it looks for if there is any factor "U" in the anaphor. If there is "U" that is considered to be matched with any of those factors other than the Gender "C". For example:

রিমা তার চাবি হারিয়ে ফেলেছে।

Here the factors of this sentence are given below.

Table 5. Factors used in the method for the sentence

শব্দ	Number	Person	Status	Gender	Parts of speech
রিমা	S	TP	N	F	NO
তার	S	TP	N	U	PN
চাবি	S	UP	U	C	NO
হারিয়ে	U	UP	U	U	V
ফেলেছে	U	TP	N	U	V

In the previous example, the pronoun "তার" matches with the noun "রিমা" in every factor except gender. The gender of the pronoun is "U". So the gender "U" can be matched with "F". So "তার" refers back to "রিমা". In this process the verbs "হারিয়ে" and "ফেলেছে" also refers back to "রিমা".

5. RESULT ANALYSIS

According to the proposed methods in case of these sentences the results were just like as hoped.

Table 6. Accuracy and error of the methods

Method	Accuracy%	Error%
Method one	20.5	79.5
Method two	38.7	61.3
Method three	38.5	61.5
Method four	43.33	56.67
Method five	76.47	23.53

Here is a table showing how many tagging of the anaphors are matched with the antecedents.

Table 7. Matching tags of the factors for the correct answers

পদ	Total correct	Number matched	Person matched	Status matched	Gender matched
সর্বনাম	33	33	33	33	4
ক্রিয়া	45	0	36	36	0

From the analysis given in the table above, it can be seen that in the correct answers of this system in case of "Pronoun" as anaphor all the number, person and status matches with the reference antecedent every time. The gender matches a very few times.

In case of "Verbal Phrases" as anaphor number and gender do not match at all. Person and status matches almost every time. The statistics given in Table 7 is based on the second data set. From 102 Anaphors there were 78 correct antecedents. The other 24 nouns which were detected as antecedents of course matched with the anaphor in almost every factor. But in real life application those are not correct. This indicates that there must be more factors or issues that are not yet determined.

6. CONCLUSION

Different methods have been tried to resolve anaphora resolution and the last method gives the best result. Although, in Bangla, tagged corpus and data set are not largely available. So, the data set needs to be manually tagged. A large corpus is essential for making more improvements. This work can be an outline for further research in anaphora resolution in Bangla.

7. REFERENCES

[1] Sikdar, Utpal Kumar, Asif Ekbal, Sriparna Saha, Olga Uryupina, and Massimo Poesio. "Anaphora Resolution for Bengali: An Experiment with Domain Adaptation", Computación y Sistemas 17, no. 2, 2013 : 137-146.

[2] Adam Meyers, "Reference Resolution", New York University, Computational Linguistics, Lecture 8, 2011-2012.

[3] Libuše Dušková , "Texture and structure. Anaphora cataphora, endophora.Cohesive ties – typology", Studies in the English Language Part 2, 1999, Chapter 36.

[4] Patrick Sturt, "The time-course of the application of binding constraintsin reference resolution", Department of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QB, UK, Received 9 January 2002; revision received 5 August 2003.

[5] Michael Strube, Stefan Rapp and Christoph Muller, "The Influence of Minimum Edit Distance on Reference Resolution", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 312-319..

[6] Sanjay Chatterji, Arnab Dhar, Biswanath Barik, Sudeshna Sarkar and Anupam Basu, "Anaphora resolution for bengali, hindi, and tamil using random tree algorithm in weka." In Proceedings of the ICON-2011, 2011.

[7] Senapati, Apurbalal, and Utpal Garain, "GuiTAR-based pronominal anaphora resolution in bengali", ACL (2), 2013.

[8] Deoskar,Tejaswini. "Techniques for anaphora resolution: A survey", Computer Science. Cornell University, 2004.

[9] Imran Q. Sayed , "Issues in anaphora resolution", The Stanford NLP. unpublished.

[10] JoseL.Vicedo and AntonioFerrandez, "Importance of Pronominal Anaphora resolution in Question Answering systems", Departamento deLenguajesySistemas Informaticos, Universidadde Alicante, Apartado 99.03080 Alicante, Spain.