Anaphora Resolution in Bangla using global discourse knowledge

Apurbalal Senapati
Computer Vision & Pattern Recognition Unit
Indian Statistical Institute
203, B.T. Road, Kolkata 700108, India
apurbalal.senapati@gmail.com

Utpal Garain
Computer Vision & Pattern Recognition Unit
Indian Statistical Institute
203, B.T. Road, Kolkata 700108, India
utpal@isical.ac.in

***Abstract*— This paper presents a pronominal anaphora resolution (PAR) approach that makes use of the global discourse knowledge along with other traditional features. So far the features used in finding the referent of an anaphoric pronoun are computed locally. Normally the sentence containing the anaphor and a few sentences immediately before form the local context. In this process, the knowledge base gets updated as more and more of the discourse is processed. Keeping this approach as the core, the present paper explores use of some prior knowledge after examining the entire discourse (whole article). Addition of this processing step improves the PAR's efficiency. This improvement is demonstrated using ICON 2011 Bangla dataset.**

Keywords: Anaphora resolution, Pronouns, Bangla, Rule-based system, ICON 2011, Evaluation.

I. INTRODUCTION

Research in anaphora resolution has progressed a lot in the last few years and several approaches have been proposed [1, 2]. Most of these works concentrate on English and because of grammatical variations and resource limitations extension of these studies to other languages was not well explored. Very recently anaphora resolution for Indic languages received attention [3]. We participated in the ICON 2011 NLP tool contest and described a pronominal anaphora resolution method for Bangla. The method is philosophically slightly different from the other existing methods. The essence of our ICON 2011 system [4] was based on the fact that a noun phrase, in a language, cannot be referred by any pronoun of the language. Rather, the set of pronouns which a noun phrase can refer is limited and varies from a noun phrase to another. Therefore, if noun phrases in a discourse go on emitting their permissible pronouns, finding referents of pronouns as encountered later in the discourse becomes easy and sometimes without using any linguistic knowledge.

This paper refines this approach by using a global discourse knowledge. In this process, the entire article is first parsed to compute certain prior knowledge which are then used during pronominal anaphora resolution. Experiment shows that addition of this approach enhance the performance of the previous system and this improvement is statistically significant. The rest of the paper briefly presents our previous system and then describes the present modification. The proposed method is illustrated with examples and its evaluation is done on ICON 2011 data.

II. ANAPHORA RESOLUTION SYSTEM FOR BANGLA BY PRONOUN EMITTING APPROACH

On encountering a pronoun, most of the existing methods first find the set of possible antecedents from the current line and a few lines backward from the pronoun (anaphor) and then filter the set of antecedents by using some rules (like agreements in number, gender, honorificity, etc.) [5 - 8]. Antecedents in many approaches are assigned weights based on some constraints/factors and then the antecedent having the highest weights is chosen as the referent [9, 10]. Our pronoun emitting approach is the reverse of almost all the existing approaches in a sense that it starts processing on encountering the possible antecedents rather than waiting for encountering a pronoun. The components (linguistic resource, antecedent object, rule base, conflict resolution system) used in this system are described below.

A. Linguistic Resource

It is the classification of pronouns based on their compatibility with noun together with some agreements like number, animate/inanimate, honorificity, etc. Table-I shows this information. It also contains some other information like honorific context {ভদ্রলোক, বাবু, ডঃ, মহাশয়,.....}, nominal relation {মা, বাবা, ভাই, বোন, দাদা, দিদি, কাকা, কাকি, জায়া, জননী, বর, কনেকে, পত্নী, বর, কনে,..}, common noun antecedent, etc.

TABLE III. LINGUISTIC RESOURCE

Category	Permissible Pronoun
Honorific Singular	তাঁর,তাঁকে,তিনি,তাঁরই,তিনিই,যাঁর,উনি,আপনি,ওঁকে,ওঁর,তিনিও,আপনারই,ইনি,আপনার…
Honorific Plural	তাঁরা,তাঁরাই,যাঁরা,উনারা,আপনারা,ওঁরা,ওঁরা,তাঁদের,...
First Person	আমি,আমরা,আমাকে,আমাদের,আমাদেরকে,আমার,মোরা,মোদের,মোর, …
Second Person	তুই,তোরা,তুমি,তোমরা,আপনি,আপনারা,তোকে,তোদের,তোমাকে,তোমাদের,আপনাকে, আপনাদের,তোর,...
Third Person	এ,এরা,এর,ও,ওরা,সে,তারা,তারও,তার,তাদের,ইনি,এনারা, উনি,ওনারা,তিনি, তেনারা, ওর,একে,এদের,ওকে, ওদের, তাকে,…
Inanimate Singular	এটা,ওটা,সেটা,এটাতে,ওটাতে,সেটাতে,এটার,ওটাতে,সেটার,এই,তা,সেটি,…
Inanimate Plural	এগুলো,ওগুলো,সেগুলো,এগুলোতে,ওগুলোতে,সেগুলোতে,এগুলোর,ওগুলোতে,সেগুলোর,সবাইয়ের,ওগুলো,এগুলির,…
Locative Pronoun	সেখানে,এখানে,এখানকার,অন্যত্র,ওখানে,সেইখানেই,ওখানেই,কোথাও,এখানটায়,সেইখান,যেখানেই,যেখানে,…
Nominal relations	মা, বাবা, ভাই, বোন, দাদা, দিদি, কাকা, কাকি, জায়া, জননী, বর, কনেকে, পত্নী, বর, কনে, …
……	……..

978-0-7695-4886-9/12 $26.00 © 2012 IEEE
DOI 10.1109/IALP.2012.25

B. Antecedent Object

- A proper noun together with permissible pronoun list (available from the *linguistic resource*) is considered as the antecedent object. It also contains other information such as sentence number, token (individual word in the text) number, honorific (if available), co-reference information (if available), etc.
- Once a pronoun is resolved, it is replaced with its antecedent and is considered as an antecedent object in subsequent processing.
- Quantifier antecedent: All nouns qualified by some quantifiers (QC, QF and QO) are considered as an antecedent objects, e.g. 'তিনটে হাতী'.
- Some common nouns (the list of these nouns is prepared from the data) are used as antecedent objects.

C. Antecedent Object List

It is a list that contains the antecedent objects as they are encountered while processing the discourse. This list maintains a LIFO (last in first out) structure.

D. Rulebase

The rule base is defined a set of heuristic rules [11, 12].

- *Honorific agreement*: In Bangla, some qualifiers are used with names to indicate a person is honorable, e.g. ভদ্রলোক, ডঃ, মহাশয়, etc. In this case, we treat that person as honorable person for future resolution. An honorable pronoun always co-refers with honorable person. Example: লতাদি বলেছে যে উনি কলকাতায় নেই। The pronoun উনি doesn't co-refer to লতাদি because the violation of honorific agreement. **Rule**: Honorific anaphora co-refers with honorable person.
- *Reflexive pronoun* (নিজ): A reflexive pronoun in Bangla always co-refers with animate occurring in the same sentence. In a simple sentence the নিজ pronoun co-refers to its subject. Example: রাম নিজেই সব কাজ করে। The pronoun নিজেই co-refers to রাম. **Rule**: The referents of reflexive pronouns are animate and co-refer to their subjects.
- *Consecutive pronouns*: In case of consecutive co-referring pronouns appearing in a sentence are person compatible (i.e. either both are in first or second person) and they also follow the honorific agreement. Example: আমি আমার সব কাজ নিজে করি । [The consecutive pronouns আমি and আমার are both in first person and hence co-refer] তুমি তোমার কাজ কর । [Pronouns তুমি and তোমার are both in second person and hence co-refer]. আমি তোমাকে ভালবাসি ।[Pronouns আমি is in first person but তোমাকে is in the second person and hence not co-refers]. However, when the pronouns are in third person, this rule is a necessary condition for treating them as co-referring. **Rule**: The person compatibility is a necessary condition for co-referring of consecutive pronouns.
- *Co-occurring pronouns*: In Bangla, some pronoun pairs (যখন/তখন, যাদের/তাদের, যার/তার, যারা/তারা) occur as co-referring in the same sentence. Example: বিশ্বভারতীর এই বিদেশী অধ্যাপক যখন আসেন, তখন তাঁর বয়স প্রায় ষাট । [যখন and তখন are co-referring]. **Rule**: The pronoun pairs (যখন/তখন, যাদের/তাদের, যার/তার, যারা/তারা) are co-referring when they appear in the same sentence.
- *Plural pronoun* (তারা/তাদের): An organization or community may co-refer to plural pronoun. Example: আই.এস.আই এখন তাদের পুরানো প্রশ্নপএ দিয়ে দেয় । [The plural pronoun তাদের co-refers to the organization আই.এস.আই].....কৃষক, যারা মূলত আদিবাসী । [The plural pronoun যারা co-refers to the community কৃষক]. **Rule**: The plural pronouns (তারা/তাদের) are permissible pronoun of an organization or community.

E. Conflict Resolution System

During anaphora resolution if there is more than one candidate referent, the resolution is done by a conflict resolution method. This module uses a set of agreement/constraints which are used in the following order: (i) pronouns are number compatible, (ii) pronouns are honorific compatible, and (iii) pronouns are person compatible. If it is still not resolved then we choose the most recent one from the antecedent list as the referent.

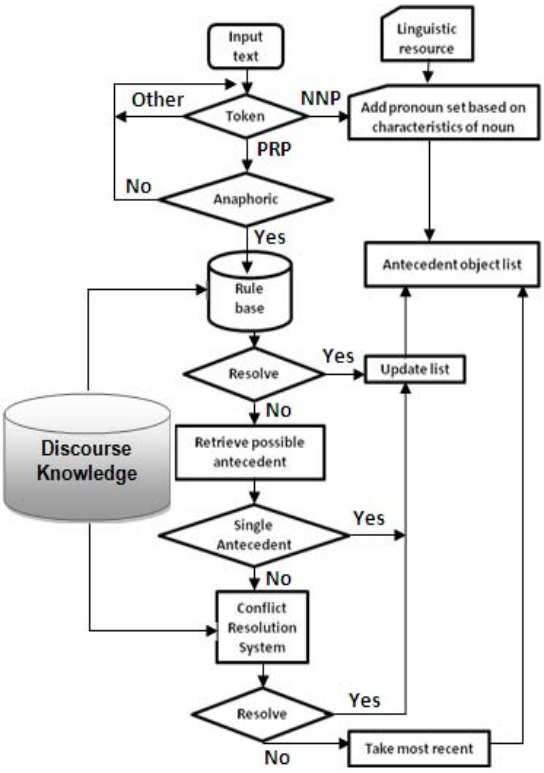

Figure 1: The architecture of the described system.

The architecture of the above system is shown in Fig-1 (consider the part excluding the one labeled as *Discourse Knowledge*). It maintains an antecedent list. When a pronoun (PRP) is found, it, at first, goes through the rule base. If the rule base resolves the PRP, then the resolved PRP is also treated as an antecedent object (by replacing the PRP with its antecedent) and the list of antecedents is updated (the antecedent is added to the list). If the rule base cannot resolve the PRP, then it finds those antecedent objects which contain this PRP in their respective pronoun lists. We consider two previous sentences for finding the candidate antecedents. If a single antecedent is found then the PRP is resolved and the antecedent list is updated accordingly. In case of more than one antecedent object then the resolution is done by the conflict resolution module.

III. THE DISCOURSE KNOWLEDGE

The discourse knowledge has been built mainly using the proper nouns occurring in the discourse with some attributes (retrieved as explained below) from the entire text.

- *Honorific Identification*: Honorificity plays an important role in anaphora resolution in Bangla. There are different degrees of honorificity in personal pronouns (Example: তুই/Tui, তুমি/Tumi, তিনি/Tini). Based on this agreement we can choose the correct antecedent. But in many cases the honorific information may be detected later in the discourse [Example: ইন্দ্রনাথকে এমন করে বলতে একমাত্র এলাই পারে। তবু তার পক্ষেও বলা সহজ নয়, তাই অস্বাভাবিক জোর লাগল গলায়। In this sentence the pronoun তার may refer to either ইন্দ্রনাথ or এলা (in the previous approach[4]) but if we know in advance that ইন্দ্রনাথ is a honorable person then তার will be resolved unambiguously. *Computational aspect*: The honorificity information can be used in two ways: checking (i) the honorific qualifier occurring immediately before the name and (ii) the honorificity of the verb [suffixed with ন]. In Bangla, a list of qualifiers {ভদ্রলোক, বাবু, ডঃ, মহাশয়, বাবু, বাবুমশাই, অধ্যাপক, মিষ্টার…} is used to specify the honorificity. Any name followed by any one of these qualifiers is an honorable person. Example: সুলেখা একবার অধ্যাপক আর.সি.ঘোষের কথাও ভাবলো। Here আর.সি.ঘোষের is an honorable person. Also a person followed by honorific verb is also an honorable person. Example: বিশ্ব শান্তির ক্ষেত্রে ভারতের সমন্বয়মুখী সাধনা ও ত্যাগের উল্লেখ করে স্টেন কোনো বলেন .... Here স্টেন কোনো is an honorable person. In Bangla, a name suffixed with 'বাবু', 'বাবুমশাই', etc. is also treated as an honorable person.
- *Hidden Person Identification*: In many cases some person appears in the discourse indirectly and they appear as an antecedent. For example, in the story *Lalu* [written by *Sarat Chandra Chottopadhay*] "শুনে লালুর বাবা বললেন, না। তাঁর নিজের কখনো মাস্টার ছিল না ...." here the person "লালুর বাবা" appears in this story five times as "লালুর বাবা" who is a separate person from লালু and doesn't exist by any other name. In the previous approach, লালু was treated as an antecedent object and the method tries to refer তাঁর with লালু. If we identify "লালুর বাবা" as a separate person (who is honorable) in advance then we can resolve তাঁর correctly. *Computational aspect*: This is identified by a rule that captures name of a person ends with case marker র/এর and followed by a nominal relation (this comes from the linguistic resource).
- *Identify the same name in different forms*: This may be of two types. The first one refers to the case where a person's name is mentioned once and then in subsequent discourse he/she is referred by his/her surname or some qualifier. For example: *Dr. Alex Anaphora is the professor of Indian Statistical Institute since 1999. Prof. Anaphora received a prestigious fellowship for the year 2011*. Here Prof. Anaphora and Dr. Alex Anaphora is the same person. This may appear in other forms like Mr. Anaphora, Dr. Anaphora etc. and it happens in many languages. The second type refers to the fact that a person may be addressed by names with different form of utterances. For example, in the story *Lalu* the person লালু appears as লেলো, লালু [similarly, রাম may appear as রাম, রামু, রামা, রেমো, David may appear as Dave, David, etc.] and if we identify these forms are identical then their honorific property will also be same. This information helps us to resolve anaphora. *Computational aspect*: For Bangla, the first type is identified by a string matching approach whereas the second type is captured by observing the differences in using vowel modifiers. We split names character wise and then excluding vowel modifiers match them character wise in same order. Example: [… রাম, লেলো, রামু, রামা, লালু, রেমো, …], first split each name character wise i.e. রাম = {র া ম}, লেলো = { ল ে ল ো }, রামু = {র া ম ু}, রামা = {র া ম া}, লালু = {ল া ল ু}, রেমো = {র ে ম ো}. Using the above rule it groups {লেলো, লালু } and {রাম, রামু, রামা, রেমো}.
- *Quantifier antecedent*: When nouns are qualified by some quantifiers, the nouns along with the quantifiers are treated as antecedent objects. *Computational aspect*: This is done by chunking. For example, তিনটে_QC হাতী_NN results in 'তিনটে হাতী' (QC_NN) as an antecedent.

Plugging of the discourse knowledge in the Previous anaphora resolution system: From the above discussion regarding discourse knowledge we have seen that this knowledge can mainly be used to resolve conflicts among the possible antecedents. It also gives some valuable information about the discourse. Hence, this knowledge is best suited to plug into the conflict resolution module as well as with the rule base (because some rules use the honorific constraints). The architecture of the above system after plugging in the discourse knowledge is shown in the Fig-1 (including the discourse knowledge component).

IV. EVALUATION

A. Data and data format

To evaluate the above approach the data provided by from ICON-2011[3] has been used. They provided the annotated data (POS tagging, chunking and name entity tagged) for five Indian languages including Bangla. The annotated data is represented by a column format. Eight types of information are presented in eight columns. The sample data is given below and detail data format is described in TABLE II.

```
#begin document (story2.txt); part 000
story2.txt   0   0  বাঁকের  NN    B-NP   o   -
story2.txt   0   1  মোড়    NN    B-NP   o   (1)
story2.txt   0   2  দেখা    VM    B-VGF  o   -
story2.txt   0   3  যাচ্ছে  VAUX  I-VGF  o   -
story2.txt   0   4  ।       SYM   I-VGF  o   -
.....................................
#end document
```

TABLE II. DESCRIPTION OF DATA FORMAT

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

B. RESULTS

The previous approach (i.e. without using global discourse knowledge) has been evaluated in the NLP tool contest organized during ICON-2011[3] and result is (in column one in TABLE III). Result in the second (in TABLE III) shows performance after plugging in the discourse knowledge and testing it on the same ICON-2011 data." Note that using the global discourse knowledge we got some extra information compared to previous approach and hence many pronominal anaphors are resolved which remained unresolved in the previous system.

V. CONCLUSION

This paper shows that use of global discourse knowledge in anaphora resolution system help in improving the efficiency of the system. The way we have extracted the information from the whole discourse is quite general in nature. So though this has been implemented in Bangla this experiment can be easily extended to other languages. Note that use of such knowledge has not been explored before for other languages including English.

TABLE III. COMPARISON RESULT

Metric		Without global discourse knowledge	With global discourse knowledge
MUC	R	11.39	48.87
	P	29.80	46.18
	F1	16.48	47.49
BCUBE	R	62.79	67.81
	P	90.52	83.24
	F1	71.15	74.74
CEAFM	R	61.35	65.39
	P	60.63	64.63
	F1	60.99	65.01
CEAFE	R	68.71	79.74
	P	44.32	64.23
	F1	53.88	71.15
BLANC	R	15.18	54.22
	P	42.85	67.81
	F1	22.42	60.26

REFERENCES

[1] R. Mitkov, "Introduction: Special Issue on Anaphora Resolution in Machine Translation and Multilingual NLP," In Machine Translation, 14(3-4): 159-161, 1999.

[2] "Modeling Unrestricted Coreference in OntoNotes," Shared task in CoNLL 2011, Portland, Oregon, USA.

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

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

[5] J. Hobbs, "Coherence and coreference," In Cognitive Science, vol. 3, pp. 67–90, 1979.

[6] B.J. Grosz, S. Weinstein and A.K. Joshi, "Centering: A Framework for Modeling the Local Coherence of Discourse," In Computational Linguistics, vol. 21, pp. 203-225, 1995.

[7] B. Baldwin, "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, 1997.

[8] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning, "A Multi-Pass Sieve for Coreference Resolution," In EMNLP, 2010.

[9] Lappin and H.J. Leass, "An algorithm for pronominal anaphora resolution," Computational Linguistics, 20(4), 535-561, 1994.

[10] A. Dhar and U. Garain, "A method for pronominal anaphora resolution in Bengali," În 6th Int. Conf. on Natural Language Processing (ICON), Student Competition section, Pune, India, 2008.

[11] A. Majumdar, "Studies in the Anaphoric Relations in Bengali," Publisher: Subarnarekha, India, 2000.

[12] G. Sengupta, "Lexical anaphors and pronouns in Bnagla," in 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), Publisher: Mouton de Gruyter, Berlin, New York, 2000.