Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analysis of Anaphora, Cataphora and Exaphora for Demonstrative Pronoun of Hindi

Pardeep Singh*, Kamlesh Dutta
CSED, National Institute of Technology
Hamirpur, Himachal Pradesh, India

Abstract*—Co-reference resolution is one of those issues of paramount importance that are required be addressed for fool proof translation and other application in which co-reference is used, especially, when target language is a free word order language. A number of algorithms have been devised to resolve co-reference resolution. The accuracy of co-reference resolution (anaphora + cataphora + exaphora) algorithm depends on the percentage of all these types of references in corpus. In order to determine its viability, we have conducted a study on Hindi corpus to find the percentage of sentences of anaphora and cataphora. Percentage of these types of sentences pave the way to evaluate the accuracy of co-reference resolution algorithms which can resolve either anaphora, cataphora or exaphora. This will guide the researchers to evaluate the parameters to consider relevance of these phoric types in study. In this paper we have considered one of the features proposed by Botley, which has been engaged to resolve the anaphora in Hindi. This tag has been analysed empirically and its values tested for a corpus. We analysed 165 news items of Ranchi Express from EMILEE corpus of plain text, wherein we exploited tag set proposed by different authors. Three values are considered of this tag. In this study there are twelve files of Ranchi Express, and have 1515 sentences from monologue. Eight files of dialogue of EMILEE corpus having 877 sentences. It is evident from the study that hundred percent accuracy of co-reference resolution is not attainable until anaphora, cataphora and exaphora is addressed to.

***Keywords*— Coreference resolution, anaphora resolution, cataphora, exaphora, annotation.**

I. INTRODUCTION

Anaphora is a referring expression that refers to some entity which have antecedently been introduced into the discourse. The process of binding (mapping) the 'referring expression' to the correct antecedent, in the given discourse, is called anaphora resolution. According to Halliday and Hassan [14], anaphora is "the cohesion (presupposition) which points back to some previous item." Hirst, more formally defines anaphora as "a device for making an abbreviated reference (containing fewer bits of disambiguating information, rather than being lexically or phonetically shorter) to some entity (or entities) in the expectation that the receiver of the discourse will be able to disabbreviate the reference and, thereby, determining the identity of the entity." Here, reference is an anaphora and the entity to whom it refers is an antecedent.

Example 1:

S1: The $Mangos_i$ were given to the $children_j$ because $they_{*i/j}$ were hungry.

S2: The $Mangos_i$ were given to the $children_j$ because $they_{*i/j}$ were ripe.

S3: The $Mangos_i$ were given to the $children_j$ because $they_{*i/j}$ were there.

Whom does each of the word "they" refer to in the example 1?

With the hearer set of belief in S1, the pronoun "they" refer to the "children" because "hungry" is the property of "they" pronoun which matched with "children". Only the "children" can feel hunger. Being the similar property of "hunger" in the pronoun "they" and noun "children" binds together as a pair of referent and referring expression.

In the same way S2 and S3 resolves the referent and the referring expression pairs by using some semantic or syntactic knowledge of sentence/ words. The problem arises when it is resolved with the help of a machine.

Example 2:

John had to go to a meeting so he decided to have a shave.

"He" refers to "John" and "John" come before "He". John is an anaphor.

In free word order language, a number of meaningful sentences can be made by scrambling the words of one sentence. This scrambling will convey the same meaning, however the 'referring expression' may change its position with the corresponding referent. 'Referring expression' found to be on the left side of referent is called anaphora and when it is found placed on the right side of the referent is called cataphora. If 'referring expression' is missing in discourse or implicit, is called exaphora.

II. COREFERENCE RESOLUTION

Coreference occurs when multiple expressions in a sentence or document refers to the same thing; or in linguistic jargon, they have the same referent. For example, in the sentence:

Radha said she would help me.

"she" and "Radha" are most likely referring to the same person or group, in which case they are co-referent. Similarly, in "I saw Raj yesterday. He was fishing by the lake". "Raj" and "he" are most likely co-referent.

The following is a sample text in Hindi, of the Article 1 of the Universal Declaration of Human Rights (by the United Nations):

Example 3:

अनुच्छेद 1 — 'सभी मनुष्यों'i को गौरव और अधिकारों के मामले में जन्मजात स्वतन्त्रता और समानता प्राप्त है। 'उन्हें'i बुद्धि और अन्तरात्मा की देन प्राप्त है और परस्पर 'उन्हें'i भाईचारे के भाव से बर्ताव करना चाहिये।

Gloss (word-to-word):

Article 1 — 'All human-beings'i to dignity and rights' matter in from-birth freedom and equality acquired is. 'Them'i to reason and conscience's endowment acquired is and always 'them'i to brotherhood's spirit with behaviour to do should.

Translation (grammatical):

Article 1 — 'All human beings'i are born free and equal in dignity and rights. 'They'i are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

In example 1 'unhone' (Anaphor) refer to 'Sabhi manuṣyoṃ' (Antecedent). Translation of above sentence have one anaphor; 'They' which refer to 'All human beings' is antecedent. Antecedent in the above example consists of three words which are tough to recover.

III. LITERATURE SURVEY

Most of the earlier work done in the field of co-reference resolution is for English and other European languages. Reference [1], have done extensive work to propose a computational solution to anaphora and co-reference resolution in English [2]. Hobbs algorithm [1], Mitkov [4]; is an effective algorithm for anaphora resolution. It uses syntactic information rather than semantic information. Reference [1] algorithm depends only on a syntactic parser plus a morphological gender and number checker. For this reason, it is often used as a baseline, when evaluating new pronominal anaphora resolution algorithms. Reference [3] work is very comprehensive in the field of anaphora resolution. JavaRAP [5], and Mitkov's [4] Anaphora Resolution System are among some of the good anaphora resolution systems.

A number of annotation schemes are available for different tasks. A number of tag set is defined by different authors [7], [9-11] for English, European languages and modified for other languages like Turkish, German, Dravidian languages etc.; to create an annotated corpus. There are six features proposed to annotate demonstrative pronoun for English language [10]. The author considers the recoverability of antecedent, direction of reference, phoric type, syntactic function, and antecedent type to annotate three genre. These corpora are the American Printing House for the Blind (APHB) Corpus, the Associated Press (AP) Corpus, and the Hansard Corpus [8]. Later, few tags are suggested and adapted the annotation scheme for Hindi [10], [12]. A machine learning approach is proposed for classification of indirect anaphora and added one more tag to previous work [9]. This tag considers the semantic category. The authors proposed that despite some syntactic constraints, semantic collocation pattern is also significant feature for indirect anaphora in Hindi [6]. An annotated corpus by adopting the lexically grounded approach of the Penn Discourse Treebank (PDTB) [10], they presented a preliminary analysis of discourse connectives in a small corpus scheme. Another study was carried out for the analysis of feature of anaphora resolution [13].

IV. METHODOLOGY

A. Feature set selection

We have used EMILLE corpus. In this corpus each occurrence of demonstrative pronoun is coded in such a manner that it could be extracted. The pronoun marked as a direct or indirect, does not specify what actually distinguishes direct anaphor from the indirect. The corpus is annotated for anaphora using scheme based on [8] and customized for Hindi corpus by [9]. In this study, we have considered one feature (Table 1). The values of the feature (Direction of Reference) may have any one of the three values. The values of this feature are 'A', 'C' & 0 (zero) for anaphora, cataphora and exaphora respectively.

TABLE I FEATURE USED FOR ANALYSIS

Feature	Value1	Value2	Value3	Value4	Value5
Direction of reference	A (anaphoric)	C (cataphoric)	0 (not applicable, exophoric or deictic)	None	None

B. Direction of Reference — Anaphoric

It is directly anaphoric if table II have values like DARMN for any demonstrative pronoun.

Example:

"*Our* offense is designed to shoot lay-ups. If we can't carry on with this offense, *we* find ourselves sitting on the bench".

TABLE II VALUE OF FEATURES FOR ANAPHORIC REFERENCE

Feature	Value	Tag
Recoverability	directly recoverable	D

Direction of reference	anaphoric	A
Phoric type	referential	R
Syntactic function	modifier	M
Antecedent type	nominal	N

C. Direction of Reference — Cataphoric

Direction of reference is said to be cataphora if antecedent is found placed on the right hand side of the anaphor.

Example:

1. If you want them, there are cookies in the kitchen. ("Them" is an instance of cataphora because it refers to cookies which haven't been mentioned in the discourse prior to that point.)
2. After he received his orders, the soldier left the barracks. ("he" is also a cataphoric reference to the soldier which is mentioned later in the discourse)

TABLE III VALUES OF FEATURES FOR CATAPHORIC REFERENCE

Feature	Value	Tag
Recoverability	directly recoverable	D
Direction of reference	cataphoric	C
Phoric type	referential	R
Syntactic function	head of noun phrase	H
Antecedent type	propositional	P

D. Non recoverable Antecedent

This feature has a value zero if it is non recoverable. Antecedent may not be present in the discourse or its reference is implicit.

Example:

"Did the gardener water *those* plants?" it is quite possible that "*those*" refers back to the preceding text, to some earlier mention of those particular plants in the discussion. But it is also possible that it refers to the environment in which the dialogue is taking place — to the "context of situation", as it is called — where the plants in question are present and can be pointed to if necessary.

V. RESULT AND DISCUSSION

Pre-tagged EMILEE corpus has been considered for study. It was annotated as per Botley's annotation scheme. The values of this tag are extracted and collected under anaphoric, cataphoric and exaphoric heading.

TABLE IV. NUMBER OF INSTANCES OF DIRECTION OF REFERENCE IN CORPUS

Corpus	Anaphoric	Cataphoric	Exaphoric
Plain	1297	169	49
Dialogue	759	42	70

Table IV reveals that the majority of co-reference in Hindi language is anaphoric in both the genre (dialogue and monologue). In plain text anaphoric, cataphoric and exaphoric references are 1297, 169 & 49 respectively. The same pattern is reflected in the corpus of dialogue with values of 759, 42 & 40 as anaphoric, cataphoric and exaphoric respectively.

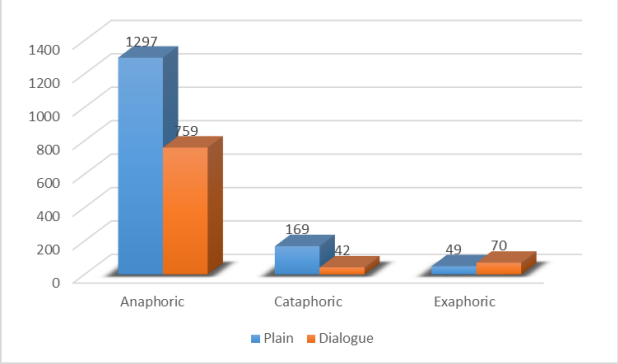

Fig. 1 Number of instances of Direction of reference in Corpus

Figure 1 shows the pictorial view of table IV. Corpus shows that all types of references in dialogue are less in number than monologue. However, the pattern of all the values of the feature remains the same.

TABLE V PERCENTAGE OF INSTANCES OF DIRECTION OF REFERENCE IN CORPUS

Corpus	Anaphoric	Cataphoric	Exaphoric
Percentage in plain corpus	85.6	11.1	3.3
Percentage in Dialogue corpus	87.1	4.8	8.1

Table V indicates that the majority of co-reference in the demonstrative pronoun in Hindi language is anaphoric in both genre (dialogue and monologue). It is found that antecedent in demonstrative pronoun can be cataphora but its value is 11 and 5 (approximately) for plain and dialogue respectively. The presence of exaphoric references are very few, however, these are difficult to recover.

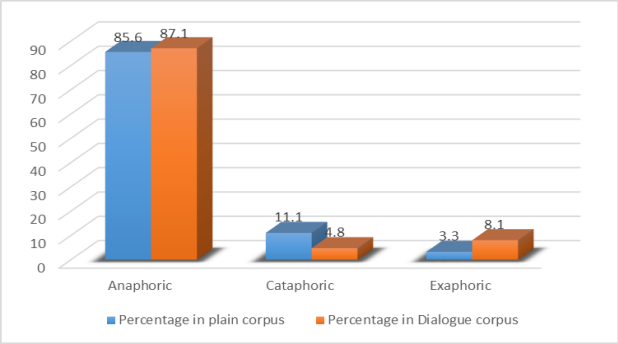

Fig. 2 Percentage of instances of Direction of reference in Corpus

In figure 2, percentage of pronoun feature 'Direction of Reference' has been shown as value of A, C and 0(zero). On axis X, values of the feature and on axis Y, the percentage of pronoun in plain text and dialogue text are shown. Figure 5 shows that 86 and 87(approximately) percentage references are anaphoric in plain and dialogue respectively. It means maximum references are anaphoric. Remaining 11 and 5 % (approximately) references are cataphoric in plain and dialogue respectively.

VI. CONCLUSION

The majority of the references, i.e, 86 percentage are anaphoric. Cataphoric and exaphoric referent sentences are 11 & 3 percentage respectively in monologue. Whereas, in dialogue corpus it has been observed that anaphoric, cataphoric and exaphoric references are there, which will have a percentage of 87, 4 & 8 respectively. It infers that exaphora has the lowest priority on the basis of its percentage. Any algorithm devised for only anaphora or cataphora cannot attain hundred percent accuracy. Co-reference resolution algorithm must address all types of references i.e. anaphoric, cataphoric and exaphoric. Exaphoric references may be left out up to a certain extent since in both types of corpuses its presence is very thin in terms of percentage. One must address the anaphora and cataphora for a substantial amount of accuracy.

REFERENCES

[1] R. Mitkov, "An integrated model for anaphora resolution," *Proceedings of the 15th conference on Computational linguistics* - Volume 2 Kyoto, Japan: Association for Computational Linguistics, 1994.

[2] R. Prasad, M. Strube, "Discourse Salience and Pronoun Resolution in Hindi," *Penn Working Papers in Linguistics*, vol 6, issue 3. UPenn, pp. 189-208, 2000.

[3] S. P. Botley, "Indirect anaphora: Testing the limits of corpus-based linguistics," *International Journal of Corpus Linguistics*, vol 1, issue 3, pp 73–112, 2006.

[4] S. Botley, A. McEnery, "Demonstratives in English: a corpus-based study," *Journal of English Linguistics*, vol. 29, pp. 7–33, 2001.

[5] K. Dutta, S. Kaushik, N. Prakash, "Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items," *The Prague Bulletin of Mathematical Linguistics* No. 95, pp 33–50, doi: 10.2478/v10108-011-0003-4, 2011.

[6] R. Prasaad, E. Miltask, A. Joshi, B. Webber, "Annotation and Data Mining of the Penn Discourse Tree Bank," *ACL Workshop on Discourse Annotation*, 2000.

[7] S. Hammami, L. H. Belguith, A. B. Hamadou, "Arabic anaphora resolution: corpora annotation with coreferential links," *The International Arab Journal of Information Technology - IAJIT* , vol. 6, no. 5, pp 480-488, 2009.

[8] S. Sinha, "A Corpus-based Account of Anaphor Resolution in Hindi," Master's thesis, University of Lancaster, UK, 2002.

[9] S. Lappin, and H. Leas, "An algorithm for pronominal anaphora resolution," *Computational Linguistics*, vol 20, issue 4, pp 535-561, 1994.

[10] P. Singh, K. Dutta, "Analysis and Comparison of Antecedent Type of Demonstrative Pronoun in Context of Co-reference Resolution: A Corpus Based Study of Hindi for Monologue and Dialogue," *Sixth IEEE International Conference on Computational Intelligence and Communication Networks (CICN 2014)*, pp 536-540, 14-16 Nov. 2014, DOI 10.1109/.122 537 DOI 10.1109/CICN.2014.122.

[11] R. Mitkov, R. Evans and C. Orasan, "*A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method*," Lecture notes in computer science, issue 2276, pp.168-186, 2002.

[12] G. Hirst, "Anaphora in Natural Language Understanding," *Springer-Verlag, Berlin*, 1981.

[13] Jerry R. Hobbs, "Pronoun Resolution," *Research Report 76-1*, Department of Computer Sciences, City College, City University of New York. August 1976.

[14] M. Halliday and R. Hasan, *"Cohesion in English,"* Longman English Language Series 9, Longman, 1976.

