Anaphora Annotation in Hindi Dependency TreeBank

Praveen Dakwale
LTRC, IIIT-H, India
dakwale.praveen@gmail.com

Himanshu Sharma
LTRC, IIIT-H, India
himanshu_s@students.iiit.ac.in

Dipti M Sharma
LTRC, IIIT-H, India
diptims@gmail.com

Abstract

In this paper, we propose a scheme for anaphora annotation in Hindi Dependency Treebank. The goal is to identify and handle the challenges that arise in the annotation of reference relations in Hindi. We identify some of the issues related to anaphora annotation specific to Hindi such as distribution of markable span, sequential annotation, representation format, annotation of multiple referents etc. The scheme hence incorporates some characteristics specific to these issues in order to achieve a consistent annotation. Most significant among these characteristics is the head-modifier separation in referent selection. The modifier-modified dependency relations inside a markable is utilized for this head-modifier distinction. A part of the Hindi Dependency Treebank, of around 2500 sentences has been annotated with anaphoric relations and an inter-annotator study was carried out which shows a significant agreement over selection of the head referent using the proposed scheme as compared to MUC annotation format. The current annotation is done for a limited set of pronominal categories.

1 Introduction

In this paper we present a scheme for annotating anaphoric relations in the Hindi Dependency TreeBank. Anaphora Resolution is one of the important problems in Natural Language Processing, and is used by various applications such as Text Summarization, Question answering etc. An anaphora annotated corpus along with other features (like POS, morph, Parse structure etc.) is required in both statistical as well as rule based anaphora resolution systems. Various corpus based studies of anaphoric variation also make use of such a corpus. While a significant number of corpora with anaphora annotation for English and other languages like Spanish, Czech etc. are available, for Indian languages, such corpora are scarce.

With a view of developing an Anaphora Resolution system in Hindi, our project aims at extending the dependency annotated (Hindi Dependency TreeBank) corpus with anaphoric relations. Hence we propose an anaphora annotation scheme in accordance with the representation format (SSF)(Bharati et al., 2007) of the Treebank, that uses attribute-value pairs to represent linguistic information. In this scheme, we attempt to address some of the issues that are commonly faced while annotating anaphora and require efficient handling. Although the scheme is developed while keeping in view the structure of the Dependency Tree-Bank, it is convertible to other formats of annotation as well.

In recent years, due to increasing interest in development of statistical systems for anaphora resolution, there have been significant attempts for creation of anaphora annotated corpora and annotation schemes. The most well known among these are MUC-7 annotation scheme (Hirschman and Chinchor, 1997) and other MUC based schemes, which are used for co-reference annotation via markup tags. The MATE/GNOME project has another important scheme suitable for different types of dialogue annotations (Poesio and Artstein, 2008). Kucova and Hajicova (2005) is also a notable work to-

26th Pacific Asia Conference on Language,Information and Computation pages 391–400

wards annotating co-reference relations in a dependency TreeBank (Czech, PragueDT). Some other proposed schemes are, in Spanish and Catalan (Recasens et al., 2007; Navarro et al., 2004) and in Basque (Aduriz et al., 2004) for 3LB corpus. A known attempt for Hindi is, for demonstrative pronouns in EMILLE corpus (Sinha, 2002).The above mentioned schemes are used for anaphora annotation in English and various other languages.

The motivation behind proposing a new scheme is that some of the challenges like annotation of distributed referent span, annotation of multiple constituents, and identification of head and modifiers are difficult to handle in above mentioned schemes. Such challenges, though faced in various languages, are more frequent in Hindi. In this paper these issues are discussed in detail and an annotation scheme is proposed in order to handle them consistently.

2 Anaphora in Hindi

A significant amount of discussion about anaphora in Hindi is available in literature. However, in this section, we discuss the categorization of anaphoric relation and pronouns in Hindi that are considered while taking decisions regarding the annotation in this project.

First, we consider classification based on pronominal forms which includes personal pronouns and reflexives as two major classes. Personal pronouns in Hindi are a separate lexical category, with the exception of first person singular and plural forms. The third person forms are also the forms of demonstrative determiners. The pronoun forms reflect the categories of person, number and respect. They include मैं(I), हम(we), तुम(you sg), आप(you resp), वह(he/she/it distal), यह(it proxml). Determiner pronouns form a major category in Hindi which include demonstratives, relatives (जो which), indefinites and interrogatives(Davison, 2003). Pronoun forms are inflected for case according to the case marking system in Hindi. It should be noted here that in Hindi gender is not directly encoded in the pronoun, however it can be accessed from verb agreement in case of nominative usage. Reflexive pronouns, which form a major pronoun category in Hindi, are not marked for gender, number or person. They include अपने - आप, स्वयं, खुद representing 'self' for different persons.

Second, we consider classification based on reference type which includes abstract and concrete reference(Dipper and Zinsmeister, 2010). Abstract reference includes the cases where an anaphor refers to an event, proposition or clause, while in concrete reference an anaphor refers to a concrete(individual) entity like noun phrase(person,place etc), quantifiers etc. It is important to note here that in Hindi same pronoun can refer to both concrete as well as abstract anaphora. For the first phase of the annotation, we consider anaphoric relations to be annotated based on the ease of identification of the referent. Thus only concrete reference type is annotated because it is easier to identify the referent in this case as compared to that in abstract anaphora. Also, we do not consider demonstratives, null pronouns, gap, ellipsis because identification of referent in these cases is relatively difficult. Reference relations can also be classified on the basis of directionality i.e. anaphora as backward reference and cataphora as forward reference. In current annotation, while anaphoric references are annotated within and across sentences, only those cataphoric pronouns are annotated which have referent in the same sentence.

3 Hindi Dependency TreeBank

The 'Hindi/Urdu Dependency Treebank' is being developed as a part of the Multi-Representational and Multi-Layered Treebank for Hindi/Urdu (Bhatt et al., 2009). It is a rich corpus with various linguistic information like POS-tag, dependency relation,morphological features in the Treebank. In order to further enrich the corpus with anaphoric reference information, we intend to annotate anaphora relations as a layer on top of the dependency layer. In the representation format of the Treebank(SSF)(Bharati et al., 2007), the information on the node is of attribute-value type, where the features are represented as values of some pre-defined attributes (e.g. name, morph, dependency relation etc.). Since Dependency relations are inherently modifier-modified type, this property can be exploited to divide the markable into head and modifiers.

4 Annotation scheme

The design of the scheme is inspired by some of the issues involved with the format of the treebank data and problems faced while using other annotation schemes. In section 4.1 we discuss some of the problems that are faced while annotating anaphora using MUC scheme, we subsequently propose the solutions to these problems that we implemented in our scheme in Section 4.2. Section 4.3 describes some additional specifications that extend the basic annotation scheme.

4.1 Design Issues

4.1.1 Markable Identification

In most of the existing schemes, the markable identification is the first step in annotation(van Deemter and Kibble, 2000). Markables are the lexical expressions, acting as potential candidates which are either referred by another referring expression or can be part of a reference chain. Without consistent specification, higher disagreement can arise among the annotators about what could constitute a markable. For instance consider example(1), in which MUC scheme would allow a markable to consist of any continuous span with arbitrary length. Thus inconsistency could arise among annotators if there is disagreement on inclusion of even a single lexical element.

(1) मैंने मोहन के भाई की किताब
I.ERG mohan.GEN brother.POSS book
ली है । मै आज उसे पढूंगा
have taken I.NOM today it.ACC will read

'I have taken Mohan's brother's book. I will read it today.'

In the above example possible markables for pronoun उसे(it) are : मोहन के भाई की किताब(Mohan's brother's book) , भाई की किताब(brother's book) and किताब(book). MUC handles this problem by considering all the above candidate markables as distinct referents, while they share common constituents. Thus there is a need to introduce an option in the scheme to represent this commonality.

4.1.2 Referent span identification

One of the most difficult problem faced while annotating anaphora is that of identifying the actual span of the referent for larger noun-phrases and named entities. This could also lead to increased disagreement in annotation because the length and content of the annotated span could differ depending on the comprehension by different annotators.

(2) राम के टूटे हुए हाथ का इलाज
ram.POSS broken hand.GEN treatment
अस्पताल में हो रहा है। उस पर
hospital.LOC be.PRS.CONT It.LOC
सोमवार तक पट्टी बंधी रहेगी।
monday till cast.NOM tie.FUT

Ram's broken hand is being treated in hospital. Cast will be tied over it till monday.

In example 2, There are 3 candidate referents of the pronoun उस पर(it) are : राम के टूटे हुए हाथ का(Ram's broken hand's) , टूटे हुए हाथ का(broken hand's) , हाथ का(hand's).
Using the MUC scheme different annotators could mark different candidates as the actual referent, thus leading to the disagreement.

However, it is much easier to identify the head of the possible referent with sufficient agreement. Also, most of the features required for anaphora resolution can be computed from the features of the head of the possible referent. For Example, in all the 3 candidates above, हाथ का (hand) is the head of the markable, and is most essential for identifying the correct referent entity.

4.1.3 Multiple Non-continuous Referents

Due to the relatively free word order of Hindi and frequent instances of gap, ellipsis, NP-coordination; cases have been observed in which there are multiple referents for a pronoun separated by intervening text-span.

(3) राम कल शाम मोहन के
Ram.NOM yesterday evening mohan.GEN
घर गया था। वे कई दिनों बाद
home went They many days after
एक दूसरे से मिले।
with each other met.

'Ram went to mohan's home yesterday evening. They met each other after many days.'

In example 3,the referent of pronoun: वे (They) includes both राम (Ram) and मोहन (Mohan).

To be able to mark the above mentioned constituents, the scheme must support annotation of multiple referents for an anaphora. However, such cases can not be handled by schemes like MUC that use simple co-indexing and marking of continuous spans.

4.1.4 Distributed referent span

In Hindi many instances are observed where the referent span is not continuous, instead, it is distributed over large distances. Such referent instances are difficult to annotate with MUC's co-indexing scheme, in which a continuous span is annotated as markable.

(4) बडा भाई कल आ रहा है मेरा ।
elder-brother tomorrow is coming my.
वह शनिवार को दिल्ली जायेगा ।
He saturday.TEMP delhi go.FUTURE .

'My elder brother is coming tomorrow. He will go to Delhi on Saturday'

In above example the referent of वह(He) is मेरा बडा भाई(my elder brother), but it is not possible to annotate it as one continuous span as used in MUC scheme.

(5) भारत की गिरती हुई अर्थव्यवस्था के लिए
India's falling economy.PURPOSE
केंद्र सरकार जिम्मेदार है । हालांकि
union-government responsible is. Though
पिछले दशक में यह काफी अच्छी स्थिति
in-last-decade it much better condition
में थी ।
in was.

'Union government is responsible for India's falling economy. Though in last decade it was in much better condition.'

Similarly, in example(5), the referent of pronoun यह(It) is भारत की अर्थव्यवस्था(India's economy) and this discontinuous referent span cannot be annotated here due to the occurrence of गिरती हुई(falling) in between.

4.1.5 Sequential annotation

Anaphors in discourse usually form chains that refer to a single entity. This evokes the issue of selection of a particular entry from the multiple previous occurrences of a single entity. The linguistic aspect of this problem addresses the issue of marking a referent that is bound to the anaphora(GB Theory). e.g. In case of reflexive, if a referent-anaphora pair occurs in a construction that inherently binds the anaphora to a particular occurrence of an entity, then it is suitable to select that occurrence as the referent. However, from a computational point of view, it is more efficient to select the nearest preceding occurrence of the entity as the referent of the anaphora because it reduces number of possible candidates for the referent of an anaphora in the previous discourse. This in turn adds to computational efficiency in anaphora resolution.

(6) जयसिंह मेवार के राजा थे ।
Jayasingh mewar.GEN king was.
वे एक महान शासक थे ।
He.NOM.HON a-great-ruler was.
उन्होंने जयपुर शहर की स्थापना की ।
He.NOM jayapur city founded.

'Jayasingh was king of mewar. He was a great ruler. He founded Jaipur city.'

In above example the referent of pronoun वे(He) in second sentence is जयसिंह(Jayasingh) in first sentence. Similarly उन्होंने(He.HON) refers to the same reference category. However, it is computationally efficient to annotate the referent of उन्होंने(He.HON) as वे(He) rather than जयसिंह(Jayasingh) since it is more nearer to उन्होंने(He.NOM), hence reducing the search space.

On the other hand consider example 7

(7) राम ने कहा कि अपनी गाडी चलाना
ram.ACC told that his car to drive
उसे पसंद है।
he.ACC likes.

'Ram told that he likes to drive his car.'

Considering sequential annotation in example 7, राम(Ram) would be selected as the referent of अपनी(his). However, reflexive pronoun अपनी(his) is bound to उसे(he.ACC), thus it would be linguistically justified to select उसे(he.ACC) as the referent.

4.1.6 Representation

Hindi Dependency TreeBank comprises of feature structures that are associated with lexical and chunk nodes. In feature structures, information(POS, morph, dependency relation etc.) is represented in

the form of attribute-value pairs. Thus, to keep the scheme consistent with the existing format, information about anaphoric relations should also be represented in the same format.

4.2 Basic Scheme Specification

4.2.1 Markable Identification

As a solution to Design Issue(Markable Identification)(Section 4.1.1), we consider chunk(Abney and Abney, 1991) to be the minimal unit of annotation. Firstly because , in Hindi dependency Treebank dependency structure has chunks[1] at node level and secondly, the features of the head element in a chunk projects its properties upto the chunk level. Chunks are already annotated with unique ids. Hence, for annotating markables, we opt to represent the markable span as a set of chunks instead of marking a continuous span. A referent span can minimally be a chunk, thus increasing the agreement by not allowing the span to be partial chunks. These chunks can later be grouped together using multiple value property. For instance, example(1) from section 4.1.1 can be chunked as follows :

(8) [$_{NP1}$ मैंने] [$_{NP2}$ मोहन के] [$_{NP3}$ भाई की]
I.ERG mohan.GEN brother.POSS
[$_{NP4}$ किताब] [$_{VGF1}$ ली है] ।[$_{NP5}$ मै]
book have taken I.NOM
[$_{NP6}$ आज] [$_{NP7}$ उसे] [$_{VGF1}$ पढूंगा ।]
today it.ACC will read

'I have taken Mohan's brother's book. I will read it today.'

In above example, one of the possible markable is मोहन के भाई की किताब(Mohan's brother's book). This can be represented as a group of 3 chunks (NP2 + NP3 + NP4).

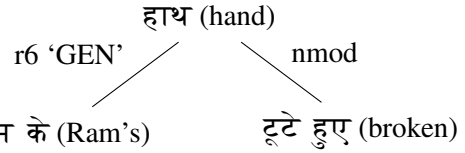
4.2.2 Reference Attributes

As discussed above in Section 4.1.2, it is easy to identify the head of the referent span as compared to the complete span. In our scheme we propose to separately annotate the easily identifiable head part (called head-referent) of the referent span and annotate the modifiers of the head-referent as a secondary information (called referent-modifiers). This could lead to a higher agreement for head-identification.
For each possible anaphora, we annotate the reference information as attribute-value pairs in the feature structure of the anaphora. Two attributes have been introduced in the feature structure namely, *'ref'* to represent the head-referent and *'refmod'* to represent the referent-modifiers. The value of these attributes specifies the unique address(es) of the above elements respectively. The addressing in current annotation is via the global address of the chunk in the document. Thus re-considering example(2) annotated with chunk information as follows :

(9) [$_{NP1}$ राम के] [$_{VGNF1}$ टूटे हुए] [$_{NP2}$ हाथ का]
ram.POSS broken hand.GEN
[$_{NP3}$ इलाज] [$_{NP4}$ अस्पताल में] [$_{VGF1}$ हो
treatment hospital.LOC be
रहा है ।] [$_{NP5}$ उस पर] [$_{NP6}$ सोमवार तक]
PRS.CONT It.LOC monday till
[$_{NP7}$ पट्टी] [$_{VGF2}$ बंधी रहेगी ।]
Caste.NOM tie.FUT

Ram's broken hand is being treated in hospital. Caste will be tied over it till monday.

The modifiers of the head of the span can be identified by looking at the dependency structure of the referent span. The dependency structure for the span राम के टूटे हुए हाथ (Ram's broken hand) would be as follows :



With the proposed scheme, if the pronoun (NP5) उस पर (It) has the referent राम के हाथ का, then it will be annotated as follows, since in this span हाथ का (NP2) is the head and राम के is the modifier :

उस पर <fs name='NP5' ref='NP2'
refmod='NP1'>

Similarly if the pronoun (NP5) उस पर (It) has the referent टूटे हुए हाथ का (broken hand), than it will be annotated as follows :

उस पर <fs name='NP5' ref='NP2'
refmod='VGNF1'>

[1]Hindi dependency treebank uses the definition of chunk as "A minimal (non recursive) phrase(partial structure) consisting of correlated,inseparable words/entities, such that the intra-chunk dependencies are not distorted"(Bharati et al., 2006)

Thus, we can see that even if different annotators identify different span for the referent, a significant agreement over the head could be achieved by separating head from the modifier.

The selection criteria for the modifiers can vary depending upon the extent of information marked and the type of problem being solved. A scheme may choose to mark only those referent-modifiers that are required to uniquely identify a referent, or it may choose to mark those referent-modifiers that help in establishing co-reference relations via lexical similarity.

4.2.3 Multiple Referents

As described in the design issues 4.1.3(Multiple Value Entries), an anaphor can have multiple head-referents. Multiple instances have been found where a part of the referent can be moved via scrambling, movement or where elements can be inserted in between. Thus it is natural to mark the referent in a way that enables maximum retrieval of information about the referent.

Chunks retain the head element feature structure and have a fixed word order internally, as is already established. Hence, by considering chunk as the minimal unit for anaphora referent annotation, it can be assured that multiple referents and their respective dependencies can be handled without any information loss. In order to annotate multiple referents, in the proposed scheme the chunk address/id of these multiple referents is specified in the '*ref*' attribute separated by a delimiter(comma). Thus reconsidering the chunked example 3 as follows :

(10) [$_{NP1}$ राम] [$_{NP2}$ कल शाम]
Ram.NOM yesterday evening
[$_{NP3}$ मोहन के] [$_{NP4}$ घर] [$_{VGF1}$ गया था।]
mohan.GEN home went
[$_{NP5}$ वे] [$_{NP6}$ कई दिनों बाद]
They many days after
[$_{NP7}$ एक दूसरे से] [$_{VGF2}$ मिले] ।
with each other met.

'Ram went to mohan's home yesterday evening. They met each other after many days.'

Thus, in above example, the feature structure of pronoun NP5(वे)(They) would be as follows:

वे <fs name='NP5' ref='NP1,NP3' refmod=''>

<ref='NP1,NP3'>implies that the pronoun has 2 head-referents, NP1 and NP3.

4.2.4 Multiple Referent-Modifiers

As discussed in section 4.1.4 (Distributed referent span), if a referent span is distributed discontiguously then it poses a problem in marking the exact span of the referent. Our scheme attempts to resolve this problem via marking the head with multiple modifiers. These modifiers are required for the correct interpretation of the pronoun; address values of all such modifier chunks are assigned in the 'refmod' attribute separated by a delimiter(/). Thus reconsidering example(4) as follows :

(11) [$_{NP1}$ बडा भाई] [$_{NP2}$ कल] [$_{VGF1}$ आ रहा है]
elder-brother tomorrow is
[$_{NP3}$ मेरा ॥] [$_{NP4}$ वह] [$_{NP5}$ शनिवार को]
coming my.He Saturday.TEMP
[$_{NP6}$ दिल्ली] [$_{VGF2}$ जायेगा ॥]
Delhi go.FUTURE .

'My elder brother is coming tomorrow. He will go to Delhi on Saturday'

In above example the referent of वह(He) is मेरा बडा भाई(my elder brother), where बडा भाई (brother)is the head and मेरा is the modifier. Hence it will be annotated as follows :

वह <fs name='NP4' ref='NP1' refmod='NP3' >

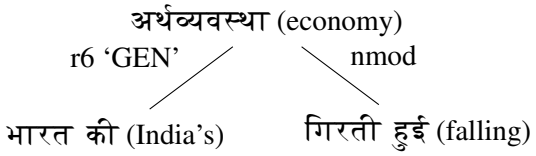
Similarly reconsidering example(5) as follows :

(12) [$_{NP1}$ भारत की] [$_{VGNF1}$ गिरती हुई]
India's falling
[$_{NP2}$ अर्थव्यवस्था के लिए] [$_{NP3}$ केंद्र सरकार]
economy.PURPOSE union-government
[$_{VGF1}$ जिम्मेदार है ॥] [$_{NP4}$ हालांकि]
is responsible. Though
[$_{NP5}$ पिछले दशक में] [$_{NP6}$ यह]
in-last-decade it
[$_{NP7}$ काफी अच्छी स्थिति में] [$_{VGF2}$ थी ॥]
in-much-better-condition was.

'Union government is responsible for India's falling economy. Though in last decade it was in much better condition.'

The referent of the pronoun NP6 (यह)(It) is (भारत की अर्थव्यवस्था)(India's economy). Head of

the span NP2 (अर्थव्यवस्था)(economy) has two modifiers NP1 (भारत की) (India's) and VGNF1 (गिरती हुई)(falling) as shown in the diagram below :



However, only NP1 is required as a modifier of NP2 for the correct interpretation of the pronoun. With the proposed scheme, we can annotate only those pronoun which are required in the referent span as shown below :

यह <fs name='NP6' ref='NP2' refmod='NP1' >

If in some case, both the modifiers are required for the interpretation of the pronoun than both the modifiers can be included in 'refmod' attribute as follows :

यह <fs name='NP6' ref='NP2'
refmod='NP1/VGNF1' >

4.2.5 Sequential annotation

In view of the computational efficiency, as discussed in section 4.1(Sequential annotation), we adopt chain marking for anaphora annotation in this scheme. That is, if an entity is referred by more than one pronouns or has repeated mentions in a discourse, then for each pronoun, we annotate the last mention of the corresponding referent-entity as the antecedent.

However, in cases where marking the nearest occurrence of the entity as referent, is not linguistically justified; the scheme allows to annotate the bound entity as the referent. Thus consider example(6) can be reconsidered as follows :

(13) [$_{NP1}$ जयसिंह] [$_{NP2}$ मेवार के] [$_{NP3}$ राजा]
Jayasinh mewar.GEN king
[$_{VGF1}$ थे ।] [$_{NP4}$ वे] [$_{NP5}$ एक महान शासक]
was. He a-great-ruler
[$_{VGF2}$ थे ।] [$_{NP6}$ उन्होंने] [$_{NP7}$ जयपुर]
was. He.NOM jayapur
[$_{NP8}$ शहर की] [$_{VGF3}$ स्थापना की ।]
city founded.

'Jayasingh was king of mewar. He was a great ruler. He founded Jaipur city.'

The referent of pronoun NP4 (वे)(He)in second sentence is NP1 (जयसिंह)(Jayasingh) in first sentence. Similarly NP6 (उन्होंने)(He) refers to the same reference category. However, it is computationally efficient to annotate the referent of NP6 as NP4 rather than NP1 since it is more nearer to NP6, hence reducing the search space. Considering sequential annotation, we annotate the pronouns NP4 and NP6 as follows

वे <fs name='NP4' ref='NP1' refmod='' >
उन्होंने <fs name='NP6' ref='NP4' refmod='' >

On the other hand consider example 7 :

(14) [$_{NP1}$ राम ने] कहा कि [$_{NP2}$ अपनी] गाडी
ram.ACC told that his car
चलाना [$_{NP3}$ उसे] पसंद है।
to drive he.ACC likes.

'Ram told that he likes to drive his car.'

Considering sequential annotation in above example, NP1(राम ने)(Ram) would be selected as the referent of NP2(अपनी)(his). However, reflexive pronoun NP2(अपनी)(his) is bound to NP3(उसे)(he.ACC), thus it would be linguistically justified to select NP3(उसे)(he.ACC) as the referent.

Hence in this example the referent of NP2(अपनी)(his) will be NP3(उसे)(he.ACC) and the referent of NP3(उसे)(he.ACC) will be NP1(राम)(Ram), with the feature structure as follows :

अपनी <fs name='NP2' ref='NP3' refmod='' >
उसे <fs name='NP3' ref='NP1' refmod='' >

4.3 Extended Scheme Specification

In this section we further describe the extended specification of the scheme that can be used to handle cases of abstract anaphora, co-reference and can be used to add additional information tags like type of anaphora, reference type, direction etc.

4.3.1 Handling Abstract Anaphora

For cases in which the referent is an event or a proposition, the main verb is marked as the referent (*'ref'*). The *'refmod'* takes the participants (modifiers) of the verb as it's values. It can either take all the participants of the event as it's values, or it can choose to take only those that are required for the

correct interpretation of the referent of the abstract anaphora.

(15) [$_{NP1}$ राम ने] [$_{NP2}$ मोहन को] [$_{NP3}$ पुरानी
Ram.ERG Mohan.DAT old
गाडी] [$_{NP4}$ ऊंचे दाम में] [$_{VGF1}$ बेची ।]
car high price-in sold
[$_{NP5}$ इससे] [$_{NP6}$ उसे] [$_{NP7}$ 5 लाख रुपए का]
Due-to-this he.DAT 5-lakh-Rs.GEN
[$_{NP8}$ लाभ] [$_{VGF2}$ हुआ ।]
profit be.PST

'Ram sold an old car to Mohan at a high price. Due to this he made a profit of 5 Lakh Rs.'

In example 6, the complete referent span is NP3+NP4+VGF1 (पुरानी गाडी ऊंचे दाम में बेची), but the head-referent is the verb VGF1 (बेची) and NP3(पुरानी गाडी), NP4(ऊंचे दाम में) are the referent-modifiers. The feature structure for pronoun NP5(इससे) is as follows :

इससे <fs name='NP5' ref='VGF1'
refmod='NP3\NP4'>

Note that only NP3 and NP4 are considered in the *'refmod'* attribute, because only these modifiers are required for the correct interpretation of the anaphoric relation.

4.3.2 Handling Co-reference

With the above scheme, the co-reference relations can also be annotated. In the case of co-reference, the value of the ref attribute would take the address/id(s) of the lexical items it co-refers with. However, including the addresses of all the lexical items (which may be large in number) can make the value field very lengthy. To avoid this, span marking is introduced. In span marking, the value contains the address of the starting and the ending lexical item joined by a delimiter(semicolon).

4.3.3 Additional Tags

Along with the reference attributes, additional tags could be incorporated in the feature structure which provide information about the anaphoric relation. Some of the important tags are :

- Pronoun Type : Personal, Reflexive, Relative, Co-relative, Indefinite.
- Referent Type : Concrete, Abstract.
- Direction : Cataphora, Anaphora.

5 Corpus Annotation and Applications

5.1 Annotation Work

In the first part of this project, 162 news items from the Treebank were considered for annotation. They contain 2477 sentences with 2122 instances of pronouns, out of which 1408 pronouns have been annotated till date. The remaining 714 pronouns were identified, but have not been annotated for the first part of annotation.

5.2 Inter-Annotator Study

We conducted Inter-Annotation studies in order to verify a higher consistency of the proposed scheme, as compared to the MUC-7 annotation framework which is commonly used for Co-reference and anaphora annotation. We divide the study in two parts as follows :

5.2.1 Experiment 1

As stated in Section 2, only Concrete reference types were annotated in the first phase of the annotation. However, in Hindi same lexical pronoun can refer to Concrete as well as Abstract reference entity and many a times it becomes difficult to identify this distinction. We first establish this by conducting an experiment which involves annotating the category of a reference type as 'Concrete','Abstract',or 'Other'(including the exo-phoric and indefinite reference types). Fleiss's Kappa (Fleiss, 1971) is used to calculate the agreement, which is a commonly used measure for calculating agreement over multiple annotators. Table 1 shows the method to interpret kappa values
The Fleiss's kappa is calculated as :

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (1)$$

The factor 1 - Pr(e) gives the degree of agreement that is attainable above chance, and, Pr(a) - Pr(e) gives the degree of agreement actually achieved above chance.

We conducted the experiment over 29 news items from the Treebank containing 446 identified pronouns across annotations by 3 raters. Annotators were asked to assign one of the three categories, as

Kappa Statistic	Strength of agreement
<0.00	Poor
0.0-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Table 1: Coefficients for the agreement-rate based on (Landis and Koch, 1977).

No. of Annotations	Agreement	Pr(a)	Pr(e)	Kappa
446	353	0.856	0.435	0.746

Table 2: Kappa statistics for Category experiment

stated above, according to the type of entity it refers to. Table 2 summarizes the experiment's results.

The non-perfect agreement for this experiment establishes that the type of the referent of a pronoun is ambiguous and hard to determine in many cases. Hence, to avoid inconsistencies in the distinction of Concrete, Abstract and Other types of reference; we separate out the concrete references in the above used data for the comparative study of the proposed scheme with MUC. We consider agreement over those pronouns in Experiment 2, for which all the annotators have a perfect agreement in concrete category.

5.2.2 Experiment 2

In the second experiment the inter-annotator analysis is conducted for the concrete pronouns separated in Experiment 1. Krippendorff's alpha (Krippendorff, 2004) was then used as a statistical measure to obtain the inter-annotator agreement. As suggested in (Passonneau, 2004) and (Poesio and Artstein, 2005) Krippendorff's alpha is a better metrics for calculating agreement for co-reference/anaphora annotation as compared to other metrics because it considers degrees of disagreement and in anaphora it is difficult to define discrete categories. Similar to (Passonneau, 2004) we consider co-reference chain as discrete categories. Experiment (2) also invloved the same data and the same raters who carries out annotation in experiment (1). Krippendorff's alpha is defined as follows :

Statistics	MUC-7	Proposed Scheme
No. of Annotations	239	239
alpha	0.825	0.880

Table 3: Krippendorff alpha agreements

$$\alpha = 1 - \frac{Do}{De} \tag{2}$$

$$Do = \frac{1}{i * c(c-1)} \sum_{i \in I} \sum_{k \in K} \sum_{k' \in K'} \mathbf{n}_{ik} \mathbf{n}_{ik'} \mathbf{d}_{kk'} \tag{3}$$

$$De = \frac{1}{i * c((i*c)-1)} \sum_{k \in K} \sum_{k' \in K'} \mathbf{n}_{k} \mathbf{n}_{k'} \mathbf{d}_{kk'} \tag{4}$$

where **I** = set of all items of annotation, **K** = set of categories, $\mathbf{n}_{ik}$ = number of times item i is given the value k, $\mathbf{n}_k$ = any number of times any item is given the value k, **i** = no. of items to be annotated, **c** = no. of annotators

The distance measure $d_{kk'}$ is defined as

$$d_{kk'} = \begin{cases} 0 & \text{if k and k' are exactly same chains} \\ 0.33 & \text{if k is a subset of k' or vice versa} \\ 0.66 & \text{if there is at least one element common between k and k'} \\ 1 & \text{if intersection of k and k' is empty} \end{cases}$$

Table 3 shows the statistics obtained for the MUC annotation and with the proposed scheme.

As shown in table 3, there is a significant increase in the Krippendorff's alpha agreement over the proposed annotation scheme, as compared to the MUC annotation scheme. This indicates that the proposed scheme with the separation of head and modifiers in the referent span helps in achieving a consistent agreement than the continuous span annotation scheme used in MUC.

5.3 Applications

The annotated data is convertible to other formats like MUC, CONLL etc. The dataset was also used for ICON-2011 Anaphora Resolution Tool Contest in Indian Languages after conversion to the required

format. A hybrid anaphora resolution system reported an average F1-score of 52.20 (ranked 1st for Hindi) using the annotated corpus for Hindi.

6 Conclusion and Future Work

In this paper we described a scheme for annotating anaphora information as a layer in Hindi Dependency Treebank. The main contribution of this paper is to discuss language specific issues that occur in anaphora annotation and outline a scheme that handles them efficiently. The identified issues relate to representation format, referent span identification etc. Decisions like sequential annotation and subtree inheritance help in reducing the computational complexity in resolution systems. The comparative inter-annotator analysis of the proposed scheme verifies that the separation of the referent span, and other features help to achieve a consistent annotation by increasing the inter-annotator agreement. The scheme can be extended for co-reference and the annotated data is convertible to other annotation formats like MUC etc.

For the purpose of this paper we have annotated concrete anaphora as described in section(2). As a further step in the project we aim to annotate abstract anaphora and co-reference relations. Also, anaphoric instances of gaps, ellipsis and demonstratives are to be included in the next phase of annotation. While the experimental results shows that proposed scheme performs well as compared to MUC format, in the future we plan to suggest improvement over MUC shceme to handle the issues discussed in this paper.

References

Steven Abney and Steven P. Abney. 1991. Parsing by chunks. In *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers.

Itziar Aduriz, Klara Ceberio, Euskal Herriko Unibertsitatea, and Daz de Ilarraza. 2004. Pronominal anaphora in basque: annotation of a real corpus. *Procesamiento del lenguaje natural*, pages 99–104.

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2006. Anncorra : Annotating corpora guidelines for pos and chunk annotation for indian languages. Technical report, LTRC, IIIT-Hyderabad.

Akshar Bharati, Rajeev Sangal, and Dipti M Sharma, 2007. *SSF: Shakti Standard Format Guide*. LTRC, IIIT-Hyderabad, India.

Rajesh Bhatt, Owen Rambow, Bhuvana Narasimhan, Dipti Misra Sharma, Martha Palmer, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*.

Alice Davison. 2003. Lexical anaphors and pronouns in hindi. In *Lexical Anaphors and Pronouns in Selected South Asian Languages: A Principled Typology*.

S. Dipper and H Zinsmeister. 2010. Towards a standard for annotating abstract anaphora. In *LREC 2010 Workshop on Language Resources and Language Technology Standards*, pages 54–59, Valletta, Malta.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Lynette Hirschman and Nancy Chinchor. 1997. Muc7 coreference task definition. In *Message Understanding Conference*.

K. Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage Publications, Inc.

Lucie Kucova and Eva Hajicova. 2005. Coreferential relations in the prague dependency treebank. In *Proceedings of the 5th International Conference on Discourse Anaphora and Anaphor Resolution*.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Borja Navarro, Ruben Izquierdo, and Maximiliano Saiz-Noeda. 2004. Exploiting semantic information for manual anaphoric annotation in cast3lb corpus. In *ACL 2004 Workshop on Discourse Annotation*.

R.J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC*, volume 4, pages 1503–1506.

M. Poesio and R. Artstein. 2005. Annotating (anaphoric) ambiguity. In *In Proc. of the Corpus Linguistics Conference*.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the arrau corpus. In *LREC*.

Marta Recasens and Maria Antnia Mart. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44:315–345.

Marta Recasens, M. Antonia Marti, and Mariona Taule. 2007. Where anaphora and coreference meet. annotation in the spanish cess-ece corpus. *Proceedings of RANLP*.

Srija Sinha. 2002. A corpus-based account of anaphor resolution in hindi. Masters thesis, University of Lancaster, UK.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*, 26:629–637.