Coreference Annotation Scheme and Relation Types for Hindi

Vandan Mujadia, Palash Gupta, Dipti Misra Sharma
FC Kohli Center on Intelligent Systems (KCIS), IIIT-H, India
vmujadia@gmail.com, palash.gupta03@gmail.com, diptims@gmail.com

Abstract

This paper describes a coreference annotation scheme, coreference annotation specific issues and their solutions through our proposed annotation scheme for Hindi. We introduce different co-reference relation types between continuous mentions of the same coreference chain such as 'Part-of', 'Function-value pair' *etc.* We used Jaccard similarity based Krippendorff's' alpha to demonstrate consistency in annotation scheme, annotation and corpora. To ease the coreference annotation process, we built a semi-automatic Coreference Annotation Tool (CAT). We also provide statistics of coreference annotation on Hindi Dependency Treebank (HDTB).

Keywords: Hindi, Coreference, Annotation Scheme

1. Introduction

There has been considerable research for coreference annotation on various languages (English(Hovy et al., 2006), French(Mitkov et al., 2000), Spanish(Recasens et al., 2007), Czech(Nedoluzhko et al., 2013), etc), on diverse domains like newspaper texts, bio-medical journals, etc. Coreference annotation is a time consuming, challenging and an expensive task. To overcome these challenges, various coreference annotation tools have been developed like CorefDraw (Harabagiu et al., 2001), GATE (Cunningham et al., 2002), PALinkA (Orasan and Sb, 2000), MMAX2 (Müller and Strube, 2001) and BART (Versley et al., 2008). All these provide text based visualization for annotation. Also there has been considerable work on coreference type relations, like, (Recasens et al., 2010) presented a typology of Near-Identity Relations and motivated the need for a middle ground category between identity and non-identity in the coreference task.

For Indian languages, co-referentially annotated corpora are few in numbers and mostly for Hindi. As stated in (Dakwale et al., 2012), most schemes which are meant to be for English are not applicable for Indian languages due to their free word order. (Dakwale et al., 2012) presents an anaphora annotation scheme and applied it on Hindi Dependency Treebank for limited set of pronominal categories, particularly for concrete anaphors. (Dakwale et al., 2012) used key-value pair attributes on anaphor chunk-head to represent its referent(s). Compared to English coreference annotation (and its representation), very little work has been done on Indian languages. This paper tries to fill that gap by describing our coreference annotation scheme for Indian Languages with Hindi examples. We also discuss issues related to coreference annotation specific to Hindi such as distribution of markable span, level of annotation, concept of chain, type relation between mentions(referential entities), full and partial membership of mentions in a chain, multi-chain membership of mentions in chains and their representation in SSF(Bharati et al., 2007). We point out several issues in the earlier annotation schemes and describe a consistent solution to those issues. Co-referentially annotated corpora were, traditionally, not annotated to capture the degree of relations between the referential entities. Our scheme also includes the degree of relation between continuous referential entities of the same chain into different relation types. This includes relations such as 'Part-of', 'Instance-of', 'Function-value pair' *etc.* These relation types would help various applications like question-answering, summarization, *etc.*

The structure of the paper is as follows. Section 2 describes possible co-referential expressions in Hindi. Section 3 describes annotation scheme and design issues and their solutions. While in sections 4, 5, 6 we discuss relation types between mentions, inter annotator agreement study and coreference annotation tool(CAT) respectively. Section 7 describe the statistics of coreference annotated corpora (Hindi dependency treebank). We conclude the paper in section 8.

2. Coreference in Hindi

When expressions refer to the same entity, they are said to be in co-referring relations. These expressions (mention or referential) can be **anaphors, nominal sequences or verb-nominal sequences**. For Hindi, we categorize anaphors according to pronominal forms and theirs reference types. As Indian languages exhibit relatively free word order, referentials (mentions) can also be, at times, fragmented. Also mentions can match a chunk, smaller (only a part) than a chunk or can be larger than a chunk. For our purpose noun sequences were divided into two types i.e. definite noun sequences and indefinite noun sequences. We took named entities, abbreviations, titles of named entities, numerals, etc in definite noun sequences. In indefinite noun sequences we included potential referential entities other than which are in definite noun sequences. For clausal or verb-clause references, we decided to annotate whole clause or sentence with all its attributes

Lexical	cref	creHead	acrefmod	acrefmodHead	crefmod	crefType	ChainHead
मोहाली$_{Mohali}$	-	-	m1%1	मोहाली$_{Mohali}$:m1	-	-	-
का$_{'s'}$	-	-	-	-	-	-	-
मोहन$_{Mohan}$	i1%1:t1	मोहन$_{Mohan}$:i1	-	-	m1		i1:t1
एक$_{one}$	i2%0:t1	-	-	-	-	-	-
अच्छा$_{nice}$	i2%0:t1	-	-	-	-	-	-
लड़का$_{boy}$	i2%1:t1	लड़का$_{boy}$:i2	-		-	noun-noun:i1	
है ।$_{is.}$							
वह$_{He}$	i3%1:t1	वह$_{He}$:i3	-	-	-	anaphora-C:i2	
मुंबई में$_{in\_Mumbai}$							
रहता है ।$_{live+PRES}$							
मोहन$_{Mohan}$	i4%1:t1	मोहन$_{Mohan}$:i4	-	-	-	noun-noun:i1	
आज $_{today}$							
हैदराबाद $_{Hyderabad}$							
आया है ।$_{came\_.}$							

Table 1: Coreference annotation for Example (1)

(1) मोहाली का मोहन एक अच्छा लड़का है । वह
Mohali's Mohan one nice boy is . He
मुंबई में रहता है । मोहन आज
in_mumbai live+PRES . Mohan today
हैदराबाद आया है ।
Hyderabad came .

'Mohali's Mohan is nice boy. He lives in Mumbai. Mohan came to Hyderabad today.'

(sentential discourse) rather than only verb as referential element.

3. Annotation Scheme

This section describes our annotation scheme and compares it with (Dakwale et al., 2012) and MUC[1] schemes.

Our coreference annotation scheme includes alltogether 7 fields. They are **cref :** This field represents the unique index for a mention, the unique index for a chain to which a mention belongs and the textual span of a mention [template- MentionId%(0/1):chainId], **crefHead :** This field represents the linguistic head of a mention, **acrefmod :** This field specifies the unique index for a modifier and its textual span, **crefmod :** This field is used to link a mention and its modifier with the unique modifier index, **crefmodHead :** This field represents the linguistic head of a modifier, **crefType :** This field specifies the type relation between mentions of the same chain, and **crefChainHead :** This field is used to mark the head mention of the chain.

Table (1) demonstrates a use-case of above mention fields using example (1). We can see in table (1) that mention indices i1, i2, i3, i4 in 'cref' tag are assigned to mentions मोहन$_{Mohan}$, एक अच्छा लड़का $_{one-nice-boy}$, वह$_{he}$ and मोहन$_{Mohan}$ respectively. Unique chain index 't1' is assigned to each mention(reference of the same concept). Mentions' textual span is specified within 'cref' tag by either %0 and %1(indicate end of mention) value. 'crefHead' indicates the linguistic head for a mention. Here at लड़का$_{boy}$, crefType='noun-noun' indicates that the mention एक अच्छा लड़का$_{one-nice-boy}$ has a 'noun-noun' relation with a mention मोहन$_{mohan}$ (coreference type relations have been discussed in next section). 'acrefmod' and 'crefmod' are used to indicate the link between the mention and its modifier, the unique index in this case is 'm1'. crefChainHead='i1:t1' is assigned on मोहन$_{mohan}$, which indicates that i1 is the head mention of the 't1' chain. Discontinuous parts of the same referential entity (single unit) captured by marking %0 and %1 with the unique mention index in 'cref' field.

We point out the gist of coreference annotation schemes ((Dakwale et al., 2012), MUC[2][1]) and theirs weaknesses, observed by us. We overcome those limitations in our scheme.

1. Our annotation scheme accommodates coreference while (Dakwale et al., 2012)'s scheme is only able to represent anaphors and their referents.
2. Our scheme operates on lexicals to include referential mentions which are may be smaller or larger than chunk/phrase while (Dakwale et al., 2012)'s' scheme operates on chunks.
3. In (Dakwale et al., 2012) scheme there is no provision to mark a relation between mentions of a chain, while MUC scheme [3][1] has provision to mark limited relation types between mentions. In our scheme we used 'crefType' field for relation marking.

[1] *http://www.nlpir.nist.gov/related_projects/muc/proceedings/co_task.html*

4. Our scheme uses indeies to represent chains while in (Dakwale et al., 2012) and MUC [4][1] schemes, We can only partially derive coreference chains with some difficulties.

5. A mention cannot be a multi-chain member in both schemes ((Dakwale et al., 2012) and MUC[5][1].

6. In both schemes ((Dakwale et al., 2012) and MUC[6][1] there is no provision to represent partial or full membership of a mention in chains.

From the example (1) (table -3), We can see that the issues mentioned above and the ones pointed out in (Dakwale et al., 2012) can resolved by our suggested scheme. 'Markable span identification' (point 2.) and 'Distributed mention(Dakwale et al., 2012)' issues can be solved by annotating coreference on each lexical of markable (see example (1), एक अच्छा लड़का$_{one-nice-boy}$). Issue 'Multiple Non-continuous referents' listed in (Dakwale et al., 2012) also can be solve by the flag value (0/1) used after % symbol in 'cref' tag, which indicates end of the lexical textual span of a mention. We can see from example (1) (table (3)) that we are capturing chain notion (points 4.-5.) by unique index (t) in cref field. We used 'crefType' field to capture the type relations between mentions (points 3.-6.).

4. Coreference relations

To capture the degree of relations between referential entities, we mark relations between the continuous mentions of the same chain. We divided relations into following broader classes. They are anaphoric relations, strong identity relations, near identity relations and weak identity relations. All-together we categorize these relations into 21 sub-categories. **Anaphoric relations** include concrete, abstract, temporal relations while **strong identity relations** include the exact lexical match between mentions. **Near identity relations** include syntactic behavior between mentions like, noun-complement, apposition, abbreviation, etc. In **weak identity relations** we include relations like part-of where one mention partially refers to other mention. For example, मुंबई पुलिस$_{mumbai-police}$ and पुलिस$_{police}$, it is clear that मुंबई पुलिस$_{mumbai-police}$ is not referring to whole पुलिस$_{police}$ but refers to sub-part of it. This deep taxonomy of relation is quite useful for question-answering and summarization alike tasks. Inferred and function-value relations are also part of weak identity relations. Following subsections explain all coreference relation types one-by-one in detail.

(2) [भाजपा नेता]$_i$ [अपने]$_i$ उग्र तेवर में
BJP_leader his.POSS fiery temper in
कुछ नहीं कहना चाहते थे ।
anything not_want_to_say+Past

BJP leader did not want to say anything in his fiery temper.

4.1. Anaphora relations :

4.1.1. Anaphora-C :

When pronominal mention has an individual or a concrete entity referent, for that we decided to mark "Anaphora-C" relation (C stands for Concrete) between those two mentions. If pronominal mention has another pronominal mention as a referent, then also we mark 'Anaphora-C' as a relation between those two pronominal mentions.
In above example (2), a pronominal mention अपने$_{his}$ is referring to concrete entity भाजपा नेता$_{BJP\_leader}$. Therefore, pronoun अपने$_{his}$ has a 'Anaphora-C' relation with भाजपा नेता$_{BJP\_leader}$.

4.1.2. Anaphora-RC :

This relation is similar to 'Anaphora-C' relation with only directional difference between a pronoun and its referent. i.e., first expression (here pronominal mention), that later co-refers with a more specific, second expression (referent mention) in the discourse. It is also known as a cataphora relation where the pronoun precedes its reference. Here 'RC' in 'Anaphora-RC' stands for Reverse Concrete.

4.1.3. Anaphora-E :

When pronominal mention has an abstract/event entity as a referent, then we decided to mark it as "Anaphora-E" relation (E stands for Event) between two mentions. If a pronominal mention has another pronoun as referent then, also we mark 'Anaphora-E' as relation between those two pronominal mentions.

(3) [ओरलैंडो इंटरनेशनल एयरपोर्ट
Orlando_International_Airport
देश – विदेश की
national_and_international.GEN
करीब साठ एयरलाइंस से जुड़ा हुआ है ।]$_i$
around_sixty_airlines being+link+pre.
[इसके]$_i$ अलावा शहर में
other-then_this other-than city_in
बस , टैक्सी और ट्रेन के अच्छे जंक्शन है ।
bus_texi_and_train's good_station be+pre

(Orlando International Airport is being link with around 60 national-international airlines)$_i$. Other-then this, city has good junctions for bus, taxi and train.

In above example (3), a pronoun इसके$_{otherthen\_this}$ which is referring to abstract entity in previous sentence. Therefore, pronominal mention इसके$_{otherthen\_this}$ has an 'Anaphora-E' relation with last sentential event.

4.1.4. Anaphora-RE :

This Anaphora-RE relation is similar to Anaphora-E relation with only directional difference. Same as 'Anaphora-RC' this type of relations are also known as cataphora.

4.1.5. Anaphora-T :

When a pronoun refers to time or time referring/representing an event or a clause in given discourse, for that we decided to mark "Anaphora-T" relation (T stands for Temporal) between those two mentions. Also, if pronominal mention has an another pronoun as a referent, then also, we decided to mark 'Anaphora-T' as a relation between those two pronouns.

(4) [मुझसे पूछा कि आप किसका नाम
i.Ablative ask+that you.SP whose name
लेना चाहेंगी ,]$_i$ [तब]$_i$ मैंने अपना नाम
like_to_take, then i.NOM i.POS name
लिया ।
take+past.

(When i was asked that whose name you would like to take)$_i$, then$_i$ I took my name

In above example (4), pronoun तब$_{then}$ is referring to time referring verb of previous sentence therefore a pronoun तब$_{then}$ has an 'Anaphora-T' relation with its previous sentence.

4.1.6. Anaphora-others :

Apart from above discussed pronominal reference types, in text there can be pronouns for which either no reference is specified or they do not have any reference. Like indefinite pronoun refers to something that is not definite or specific or exact. Indefinite pronouns include quantifiers (some, any, enough, several, many, much); universals (all, both, every, each); and partitives (any, anyone, anybody, either, neither, no, nobody, some, someone). Many of the indefinite pronouns can function as determiners. In Hindi कोई$_{someone}$, इतना$_{this\_much}$ and कुछ$_{some}$ are indefinite pronouns. In Hindi कुछ$_{some}$ is used to indicate a portion or quantity of some entity. It is also use to indicate unspecified quantity of countable entities and unspecified portion of uncountable entities. The indefinite pronoun कोई$_{someone}$ used to indicate the absence of a portion or quantity of some entity. This indefinite pronouns are marked with 'Anaphora-others' relation.

4.2. Strong Identity relation :

In this strong identity relation, we decided to mark only exact match (same lexicals/strings) and partial match (matched with entity head) mentions under 'strong identity relation' as they refer to same real world entity with same lexical patterns.

4.2.1. Coreference-Identity :

This only relation comes under strong Identity relation where two mentions have same lexical pattern and same mention head which are identical to each other. For example, सचिन रमेश तेंदुलकर$_{Sachin_Ramesh_Tendulker}$ and सचिन$_{Sachin}$,सचिन रमेश तेंदुलकर$_{Sachin_Ramesh_Tendulker}$ and तेंदुलकर$_{Tendulker}$,राजग सरकार$_{Rajag_goverment}$ and सरकार$_{Goverment}$ are pairs of mentions which are identical to each other on referential as well as lexical bases but राजग सरकार$_{Rajag_goverment}$ and यूपीए सरकार$_{UPA_goverment}$, राजग सरकार$_{Rajag_goverment}$ and राजग$_{Rajag}$ are certainly not.

4.3. Near Identity relations :

In these type of relations, two referents are representing same discourse entity although they may have different lexical pattern and also they can be in syntactic constructions. We divided this type into mainly 6 sub-types, base on their significance. Those types are discussed in following subsections.

4.3.1. Coreference-Apposition :

Apposition is a grammatical construction in which two mentions (a noun or noun phrase) are occurred subsequently, where one is serving to identify the other. Coreference-Apposition relation occurs when there is a proper noun and followed by its description, which also has an independent capability to replace previous proper noun for further reference in discourse. The mention pair having this property are said to be in apposition relation and we decided to mark this relation type between them. i.e. ,सचिन तेंदुलकर, एक महान खिलाड़ी है. (Sachin Tendulker is one of the greatest player.) In this example सचिन तेंदुलकर$_{Sachin_Tendulker}$ is also explained as एक महान खिलाड़ी$_{one\_of\_the\_gretast\_player}$, hence एक महान खिलाड़ी$_{one\_of\_the\_gretast\_player}$ is in apposition relation with सचिन तेंदुलकर$_{Sachin_Tendulker}$.

(5) [जयेंद्र सरस्वती]$_i$,
Jayandra_Saraswati ,
कांची कामकोटि पीठ के
Kanchi_Kamakoti_unioun_of
[शंकराचार्य ने]$_i$ वार्ता के जरिये
Shankaracharya.ACC through_talk
अयोध्या मसला सुलझाने की दिशा में एक बार फिर
Ayadhya_issue resolve towards once_again
मदद की पेशकश की है ।
help offer

Jayendra Saraswati, Shankaracharya of Kanchi Kamakoti union once again offers help to resolve the Ayodhya issue through talks.

In above example (5), mention जयेंद्र सरस्वती $_{Jayandra\_Saraswati}$ is a proper noun and is followed by कांची कामकोटि पीठ के शंकराचार्य (Shankaracharya of Kanchi Kamakoti union). So according to our definition, जयेंद्र सरस्वती $_{Jayandra\_Saraswati}$ has 'Coreference-Apposition' relation with कांची कामकोटि पीठ के शंकराचार्य (Shankaracharya of Kanchi Kamakoti union).

4.3.2. Coreference-NounComplement :

Like Apposition, Noun Complement is a grammatical construction in which two nouns or/and noun phrases are placed side by side, and first noun phrase uses for showing designation or position of second noun phrase. In a way this relation has exactly reverse direction then

apposition relation. A designation followed by a proper noun phrase make the Coreference-NounComplement relation. This relation is most commonly observed relation in Hindi dependency treebank .

(6) [पूर्व सलामी बल्लेबाज]$_i$
former_opening_batsman
[रमीज का]$_i$ बोर्ड में सीईओ
Rameez board_in CEO
रहते हुए टीवी कमेंटरी करने
being TV doing+commentary
की काफ़ी आलोचना हो रही थी ।
criticize+past+cont.

Former opener, Rameez was heavily criticized for doing TV commentary while being a CEO in the board.

In above example (6), two mentions, पूर्व सलामी बल्लेबाज$_{former_opener}$ and रमीज$_{Rameez}$ are placed side by side and first noun sequence is showing designation for the second one. Therefore, mention रमीज$_{Rameez}$ has a 'Coreference-NounComplement' relation with पूर्व सलामी बल्लेबाज$_{former_opener}$.

4.3.3. Coreference-Abbreviation :

Like previous relation types, this relation type does not have linguistic significant but sentential construction for this type is very frequent in text. This relation type is used to show the relation between shortened form of a noun sequence and a noun sequence. Usually, shortened form consists of a letter or group of letters taken from the initial of individual words of a noun sequence. Specially for Hindi, where there can be two types of abbreviation, 1) noun sequence is in romanize Devanagari and it is abbreviated from their initials. 2) noun sequence is in Devanagari and its initials are taken in abbreviated form.

(7) रिलायंस ने अपनी गैस [नेशनल थर्मल पॉवर प्लांट]$_i$ (
Reliance own_gas national_thermal_plant (
[एनटीपीसी]$_i$) को बेची ।
NTPC) sell+past

Reliance sold its gas to (National Thermal Power Plant)$_i$ (NTPC$_i$) .

In above example (7), नेशनल थर्मल पॉवर प्लांट$_{national_thermal_plant}$ is in romanized Devanagari and there individual word's initials are taken for abbreviation. So एनटीपीसी$_{NTPC}$ has a relation 'Coreference-Abbreviation' with नेशनल थर्मल पॉवर प्लांट$_{national_thermal_plant}$ noun sequences.

4.3.4. Coreference-RAbbreviation :

Like Coreference-RAbbreviation relation-type, this relation-type has same significant with only directional difference between mention and its referent. As 'R' indicates the reverse direction.

4.3.5. Coreference-Noun-Noun :

This relation is applicable when none of the syntactic (above mentioned), semantic (mentioned in next subsection) relations between two mentions are applicable, and still they are referring to a same discourse entity. So this relation has a significance when there are no other relation are applicable between two mentions.

(8) घटना की सूचना [प्रबंधक]$_i$ [
Incident's_information' manager
सुबोदीप चक्रवर्ती को]$_i$ दी ।
Subodeep_Chakraborty.DAT give+past
[प्रबंधक ने]$_i$ पुलिस को इस घटना से
Manager police.DAT this_incident
अवगत कराया ।
inform+past

Information about incident was given to Manager$_i$ (Subodeep Chakraborty)$_i$. Manager$_i$ informed police about the incident.

In above example (8), three mentions (with i index) are in referring to same entity, where last mention प्रबंधक$_{manager}$ is referring to second mention सुबोदीप चक्रवर्ती$_{Subodeep\_Chakraborty}$. There is no syntactic linkage between mention प्रबंधक$_{manager}$ and mention सुबोदीप चक्रवर्ती$_{Subodeep\_Chakraborty}$ thus, mention प्रबंधक$_{manager}$ has 'Coreference-Noun-Noun' relation with mention सुबोदीप चक्रवर्ती$_{Subodeep\_Chakraborty}$.

4.3.6. Coreference-Noun-Verb :

This relation 'Coreference-Noun-Verb' is applicable when from two mentions, one mention is of a noun sequence and it is referring to an event (noun-verb sequence) in discourse. In this situation, we decided to mark 'Coreference-Noun-Verb' between these two mentions.

4.4. Weak Identity relations :

In this sub type of relation, we capture those relations in which, mentions are related to each other by various semantic notions. Conceptually, we can relate them through various word-net relations, though lexically they are not identical to each other but conceptually they are related to each other. These relations are mainly useful in question-answering kind of applications.

4.4.1. Coreference-PartOf :

From the Hindi dependency treebank, we identified that a noun sequence and a pronoun can refer to multiple mentions. This reference type is most obvious and unique among the all where, one mention is physically or conceptually part of other referring mention or we can say current mention is a subset of referring mention.

(9) [राजस्थान में]$_i$ आंधी से [उत्तरी हिस्से में]$_i$
Rajasthan+loc storm northern_part+loc
पेड़ उखड़ गए ।
trees uproot+past .

Tree uprooted by the storm in the northern part of Rajasthan.

In above example (9), mention उत्तरी हिस्से $_{northern\_part}$ is part of राजस्थान$_{Rajasthan}$. It is not referring to whole entity but only some part. In this case the noun sequence उत्तरी हिस्से $_{northern\_part}$ has a 'Coreference-PartOf' relation type with राजस्थान$_{Rajasthan}$.

4.4.2. Coreference-RPartOf :

Coreference-RpartOf relation is similar to Coreference-PartOf with only directional difference. Like Coreference-PartOf, this is also a relation between two mentions, where in mention sequence, the first mention is subset of second mention or we can say that first mention is physically or conceptually part of second mention.

(10) [भारत]$_i$ और [पाकिस्तान के]$_j$
India and Pakistan
बीच बढ़ते संबंधों का असर कश्मीर के
impact_of_the_growing_ties Kashmir.GEN
किचन में भी देखा जा रहा है । लोग इसे
kitchen_in see+PP. People this
[दोनों देशों के]$_{i,j}$ बीच बढ़ते रिश्ते के
two_countries growing_relationship
परिणाम के रूप में देख रहे हैं ।
result as see+pres+cont.

The impact of the growing ties between India and Pakistan, is seen in the kitchen of Kashmir. People are seeing it as a result of growing relationship between two countries .

In example (10), दोनों देशों$_{two\_countries}$ is bigger set while भारत$_{India}$ and पाकिस्तान$_{Pakistan}$ are part/smaller set of it, thus दोनों देशों$_{two\_countries}$ has a 'Coreference-RPartOf' relation with both भारत$_{India}$ and पाकिस्तान$_{Pakistan}$.

(11) [आम]$_i$ खाने के बाद उसने [
Mango eat+cont._after he+NOM
गुठली को]$_i$ फेंक दिया ।
kernel+DAT throw+past

After eating mango$_i$, he threw kernel$_i$.

4.4.3. Coreference-Inferred :

This Inferred relation thoroughly based on different lexical and semantic relation between mentions. This lexical and semantic relation (Coreference-Inferred) includes synonymous, Hyponymy and Hypernymy, Meronymy and Holonymy (Part-whole relation), Entailment, Troponymy, Antonymy. This relation occurs when an mention can be inferred or derived from its predecessor mentions. For example, वाहन$_{vehicle}$, गाड़ी$_{car}$ both can be inferred from each other, because they share the same synset.
In example (11), गुठली$_{kernel}$ has wordnet relation with आम$_{mango}$. Thus they link each other with 'Coreference-Inferred' relation.

4.4.4. Coreference-Function-Value :

In a sentence, function or variable is a notation that specifies places, where the substitution of certain quantifiable value, may take place. This notion is related to a placeholder (a symbol that will later has some value), or a character that stands for an specified function. It is also somewhere related to copula, but here we are accounting only those functions which are capable to holding only quantifiable values. Here for example, it can be anything like संख्या$_{number}$, तापमान$_{temperature}$, क्रमांक $_{rank}$ and गुण$_{mark}$, etc.

(12) [विमान संख्या]$_i$ [आईसी – ८०३ के]$_i$
Flight_number IC-803's
कैप्टन ने बिना समय गंवाए
caption.ACC without_time_weasting
विमान को वापस दिल्ली की तरफ मोड़ दिया ।
aircraft.DAT back to_Delhi.LOC turn+past

(Flight Number)$_i$ (IC - 503)$_i$'s' caption turned the aircraft back to Delhi without losing time .

In above example (12), mention आईसी – ८०३$_{IC-803}$ is not at all a reference of विमान संख्या$_{Flight\_number}$ but they shared some attributes. That is, one is the function and other describes the value that the function can take. So आईसी – ८०३$_{IC-803}$ has relation 'Coreference-Function-Value' with विमान संख्या$_{Flight\_number}$.

5. Inter-Annotator Study

Coreference annotation is defined as a process of language corpora annotation, to indicate which textual expression have been used to co-specify the same entity in the discourse. When such an annotated corpora are collected from different coders, the reliability of the annotated data has to be quantified. Many times due to annotator's preferences, they do prejudice annotation, to standardize annotation process these types of errors have to be quantified. For coreference annotation, we used various reliability metrics to quantify the annotation scheme and annotation. As mention in (Passonneau, 2004), Krippendorff's alpha (Krippendorff, 2012) is a better metric for calculating agreement for co-reference annotation as compared to other metrics. Because it considers degree of disagreement and can be apply for more than two annotators. Similar to (Dakwale et al., 2012) as explained in (Passonneau, 2004), we also consider coreference chain as discrete categories.

Equation (Figure 1) demonstrate the Kripandorff's alpha, Where PD_o is probability of observed disagreement and PD_E is probability of expected disagree-

$$\alpha = 1 - \frac{p_{D_O}}{p_{D_E}} = 1 - \frac{rm-1}{m-1}\frac{\sum_i \sum_b \sum_{c>b} n_{b_i} n_{c_i} \delta_{bc}}{\sum_b \sum_c n_b n_c \delta_{bc}}$$

Figure 1: Kripandoff alpha

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Figure 2: Jaccard Index

ments. For r coding units and m coders, the equation calculates the agreement among the annotators by summing disagreement coefficient within and across the annotators. For every pair of values b and c (for sets), δ_{bc} is the distance between the values. n_{bi} is the number of times the value b occurred in ith unit. In nominal scales $\delta = 0$ when b = c (equivalent sets); otherwise $\delta = 1$ (different sets). The δ value in above equation is depends upon comparison of two sets.

As described in (Passonneau, 2004) the relation between two sets can be describe in four different ways i.e, identity, subscription, intersection and disjunctions and their δ values are, 0 for identity, 0.33 for subsumption, .67 for intersection and 1 for disjunctions. As discussed in (Artstein and Poesio, 2008), the assignment of δ value mislead the agreement, because it is not able to capture the length of sets (cardinality of sets). i.e. for the same sets discussed in (Passonneau, 2004), {C, H, J, K} and {C, H} has subsumption relation and {C, H, J, K} and {C} also has subsumption relation so according to (Passonneau, 2004), one needs to give 0.33 δ value. These two sets {C, H} and {C} have 2 and 1 elements respectively, and this difference in sets cannot be captured here. Therefore we use **Jaccard index** also known as the Jaccard similarity coefficient , for comparing the similarity anddiversity of sample sets, to capture this difference as mention in (Artstein and Poesio, 2008). The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets (Jaccard equation - Figure 2). We use new δ define as follow:

$$\delta = \begin{cases} 0, & \text{for } \textit{identity} \\ 0.33 * J, & \text{for } \textit{subsumption} \\ 0.67 * J, & \text{for } \textit{intersection} \\ 1 * J, & \text{for } \textit{disjunctions} \end{cases} \quad (1)$$

With the mentioned annotation scheme, we measured the inter annotator agreement score on 60 coreference annotated documents with 3 annotators on Krippendorff's alpha with Jaccard similarity index. We calculated agreement score on two different set of documents (60 each) with same annotators and got average agreement score around 87 % on 1400 co-refereeing mentions.

As the Hindi dependency treebank has several layers of annotation from part-of-speech to Karaka based dependency(Begum et al., 2008) and on the top of that annotators annotated coreference relations, links and span of mentions. Annotators can refer these layered information throw annotation tool (CAT). All these detail information tuned corpora tends to lead us on relatively higher agreement score.

6. Coreference Annotation Tool (CAT)

We built coreference annotation tool (CAT) to help and automate the complex process of coreference annotation. This web based tool has functionality to identify annotation errors and provide an assistance to annotators. Other annotation functionality like semi-automatic annotation, textual representation are implemented after concerning with the annotators. CAT mainly deals with 3 aspects of Coreference annotation 1) **Mention identification** 2) **Initial feed sets/chains generator** 3) **Automatic mention and chain head identification**.

6.1. Mention identification :

We use automatic mention identifier tool [7] and configure it in this coreference annotation tool (CAT). From where annotators can guide them self in mention selection or directly add those mentions in further annotation process.

6.2. Initial feed chain generation :

On extracted mentions, CAT also provide an option to generate initial coreference chains/sets base on several string match, named dictionary and dependency relations (optional) base rules. This is up to annotators to use this facility or not because sometimes some of the rules generate noisy chains. In this kind of tedious annotation, annotators sometimes make mistakes. i.e wrong coreference chain selection for a given mention. We try to tackle these kind of issues by using derived dictionary (from dbpedia) and dependency based rules in CAT. These rules give alert to annotators on doubtable annotations. These warning messages can be ignored. Although annotators can edit tool generated automatic annotation for smooth working.

6.3. Head Selection :

As Hindi is a head final language (Benmamoun et al., 2009), linguistic head of a noun sequence/mention is mostly the last word and, main verb for the event (verb-noun sequence) mention. We built automatic mention head identifier based on that observation. We also observed that for a coreference chain, its head/governing element lies in its first mention thus we also added an automatic chain head marking facility to CAT. Annotators can always edit/update/delete automated annotation to correct the annotation process.

[7] https://github.com/vmujadia/MentionIdentifier

7. Annotated Data

Around 9000 sentences of Hindi dependency treebank(Begum et al., 2008) have been annotated with coreference and their relations with our described annotation scheme. During these process CAT was used by annotators for assistance.

Hindi Dep. TreeBank	Size
# Documents	600
# Sentences	9000
# Tokens	90000

Table 2: Corpus detail

Mention Type	occurrences
Personal Pronouns	1200
Reflexive Pronouns	787
Relative Pronouns	345
Locative Pronouns	546
Verb-nominal sequences	500
Definite noun sequences	3000
Indefinite noun sequences	1677

Table 3: Distribution of co-referential entities

The annotation task was carried out on 600 news text documents of treebank. Table (2) shows the corpus statistics while table (3) shows the distribution of co-referential entities across the corpus.

8. Conclusion and future work

In this paper, we described a scheme for coreference annotation and applied to the Hindi Dependency Treebank. In future this scheme can be validated for other languages especially Indian languages. The main contribution of this work is to define and discuss different types of co-referential relations between the referential entities to account the need of applications like question-answering, summarization, sentiment analysis, etc. The inter annotator agreement shows that proposed scheme performs consistently well.

9. Bibliographical References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Begum, R., Husain, S., Dhwaj, A., Sharma, D. M., Bai, L., and Sangal, R. (2008). Dependency annotation scheme for indian languages. In *IJCNLP*, pages 721–726. Citeseer.

Benmamoun, E., Bhatia, A., and Polinsky, M. (2009). Closest conjunct agreement in head final languages. *Linguistic variation yearbook*, 9(1):67–88.

Bharati, A., Sangal, R., and Sharma, D. M. (2007). Ssf: Shakti standard format guide. *Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India*, pages 1–25.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust nlp tools and applications. In *ACL*, pages 168–175.

Dakwale, P., Sharma, H., and Sharma, D. M. (2012). Anaphora annotation in hindi dependency treebank. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 391–400.

Harabagiu, S. M., Bunescu, R. C., and Maiorano, S. J. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., and Sotirova, V. (2000). Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pages 49–58. Citeseer.

Müller, C. and Strube, M. (2001). Annotating anaphoric and bridging relations with mmax. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, pages 1–6. Association for Computational Linguistics.

Nedoluzhko, A., Mírovskỳ, J., and Novák, M. (2013). A coreferentially annotated corpus and anaphora resolution for czech. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference" Dialogue*.

Orasan, C. and Sb, W. W. (2000). Clinka a coreferential links annotator.

Passonneau, R. (2004). Computing reliability for coreference annotation.

Recasens, M., Marti, M. A., and Taulé, M. (2007). Where anaphora and coreference meet. annotation in the spanish cess-ece corpus. In *Proceedings of RANLP*. Citeseer.

Recasens, M., Hovy, E. H., and Martí, M. A. (2010). A typology of near-identity relations for coreference (nident). In *LREC*.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution. In *Proceedings of the 46th NAACL: Demo Session*, pages 9–12. Association for Computational Linguistics.