PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 95 APRIL 2011 33–50

Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items

Kamlesh Dutta[a], Saroj Kaushik[b], Nupur Prakash[c]

[a] National Institute of Technology, Hamirpur
[b] Indian Institute of Technology, Delhi
[c] Guru Gobind Singh Indra Prastha University

Abstract

In this paper, we present machine learning approach for the classification indirect anaphora in Hindi corpus. The direct anaphora is able to find the noun phrase antecedent within a sentence or across few sentences. On the other hand indirect anaphora does not have explicit referent in the discourse. We suggest looking for certain patterns following the indirect anaphor and marking demonstrative pronoun as directly or indirectly anaphoric accordingly. Our focus of study is pronouns without noun phrase antecedent. We analyzed 177 news items having 1334 sentences, 780 demonstrative pronouns of which 97 (12.44 %) were indirectly anaphoric. The experiment with machine learning approaches for the classification of these pronouns based on the semantic cue provided by the collocation patterns following the pronoun is also carried out.

1. Introduction

The automatic classification of indirect anaphora has attracted little attention of computational linguists. Indirect anaphora poses difficulty in designing anaphora resolution system required in various natural language applications (Mitkov, 1997) as the anaphor and antecedent do not exist explicitly in the text. Demonstrative pronouns have been found to be used as direct or indirect anaphora. For the purpose of the correct semantic interpretation of the text, it is important to be able to classify demonstrative pronouns as direct or indirect anaphora in the first instance and as-

Corresponding author: kd@nitham.ac.in
Cite as: Kamlesh Dutta, Saroj Kaushik, Nupur 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, 2011, pp. 33–50. doi: 10.2478/v10108-011-0003-4.

sign correct semantic to the demonstrative pronouns acting as indirect anaphora in the next phase. Since explicit referent for indirect anaphora does not exist in the text, such an anaphora need to be identified and semantically understood in order to automatically understand the meaning of the text. This kind of anaphora is important for natural language tasks such as discourse resolution, information extraction, machine translation and language generation.

Among the recent activities in dealing with indirect anaphora (Fan et al., 2005) is based on Semantic path whereas (Gasperin and Viera, 2004) used word similarity lists for Portugeese corpus. Gundel et al. (2005) presented encoding scheme for indirect anaphora for Santa Barbara Corpus of Spoken American English. The work of Gundel et al. (2007) is based on the hypothesis of activation and focus hypothesis for New York Times news corpus. Kerstin and S.Hansen-Schirra (2003) presented multiplayer annotation for German News Paper corpus. Gelbukh and Sidorov (1999) presented indirect anaphora resolution based on the use of a dictionary of prototypic scenarios associated with each headword, and also of a thesaurus of the standard type. Boyad et al. (2005) have demonstrated the automatic classification of "it" for non-referential properties. Each work notes that dealing automatically with indirect anaphora is still a challenging task. All theories are based on semantic or conceptual structures and therefore automating their resolution requires more efforts. However one thing about the indirect anaphora is very clear that though it is inferable from the extended text, no explicit feature allow us to assign a relationship between anaphor and antecedent. Further the amount of such anaphora is sparse and a suitable automatic classification scheme needs to be evolved as its level of resolution does affect the anaphor resolution process.

In the present paper we develop an automatic classification scheme for indirect anaphora for Hindi text, which we believe, has not been attempted so far. Hindi has large number of demonstrative pronouns, which may have a direct referent or indirect one. We shall first identify the features that could be used for prediction of demonstrative pronoun's referentiallity. We shall also perform experiments using machine-learning algorithms to have an insight into the complexity of problem so that further refinements can be carried out. According to Schwarz (2001) we do not only categorize direct anaphoric relations, in which two expressions refer to the same extra-linguistic entity. In order to include more implicit relations between text elements, we also consider relations other than referential identity to be coreferential, which we call indirect anaphoric relations. A semantic and conceptual relation rather than a grammatical or lexical one links these identities. According to Mitkov (2002) indirect anaphora can be thought of as coreference between a word and an entity implicitly introduced in the text before. This gives rise to two problems with respect to the indirect anaphora: (a) detection of indirect anaphora, and (b) assigning an appropriate antecedent which in this case not available explicitly (Gelbukh and Sidorov, 1999).

2. Indirect Anaphora in Hindi

We first give a brief description of some key grammatical aspects of the demonstrative pronominal, and then discuss the issue of anaphoricity in Hindi. A list of possible demonstrative pronouns and their indirect anaphoricity behavior is given in Table 1. As evident, the number of pronoun usage is very large. Some of the pronouns can have indirect as well as direct anaphoricity whereas others have a direct antecedent in the discourse text.

The root form of these demonstrative pronouns is "yeh", "veh", "iss", "uss", "inn", "unn", "yahaan", "vahaan", "eissa", "veissa". The case marking modifies the pronouns and indicates the relation of pronoun with the neighbouring words. The case marker is added separately and the pronoun modifies accordingly. The agreement inflection is marked for person, number, and gender. In some readings the modified pronoun appears as a single word where as in others it is represented as two separated words. "inmein" "इनमे" (in these) can be written as "in mein" "इन में" or "inmein" "इनमें". Both forms are acceptable in written Hindi. However for our study we assume the modified pronoun as a single word. Various inflections after adding case marker to root word "iss" (this/it) is shown in Table 2.

Pronouns can appear as a noun or a modifier of noun. Noun form occurrences are governed by the case marking. Pronouns appearing as a noun in ergative, dative, and accusative forms require exact antecedent in the discourse. For example ergative cases (Pandharipande and Kachru, 1977), marked with case marker, "ne", expresses actor/ agent/ subject in perfective tenses for transitive verbs, as shown in sentence (1). The perfective form is indicative of pronoun + "ne" behaving as a noun phrase and the pronoun maps to some agent in the discourse. Non-animate nouns are not marked with ergative case. Therefore, normally the pronouns with these case forms do not exhibit the indirect anaphora.

(1) उन्होंने कहा कि महिला आरक्षण में विशिष्ट वर्गों के लिए अलग से आरक्षण की मांग सही नहीं है .
Unhon-ne kahaa ki mahilaa aarakshan mein vishisht vargon ke liye alag se aarakshan kii maang sahi nahiin hei.
He/She/They said that in the women's reservation demand for separate reservation for special category is not right.

On the other hand, several other forms of pronoun act as a modifier of noun and perfectly behave as a demonstrative pronoun. Such pronouns may be indirectly anaphoric as shown in sentence (2).

Pronoun in Hindi	Roman Gloss	English Pronoun	Indirect Anaphora
यह	yeh	this/it	yes
वह	veh	that	no
ये	ye	these	no
वे	ve	they	no
इस	iss	this/it	yes
इसे	isse	it	yes
इसी	isii	this	yes
उसी	usii	that	yes
इसका	isska	its	yes
इसकी	isskii	its	yes
इसके	isske	its	no
इसने	issne	it	no
इससे	iss-se	with it	no
इसमें	iss-mein	in it	yes
उस	uss	him/he/itr	no
उसे	usse	him/her/it	no
उसका	uss-ka	his/her/its	no
उसके	uss-ke	his/her/its	no
उसमें	uss-mein	in it	no
उसकी	uss-kii	his/her/its	no
उसने	uss-ne	he/she	no
उससे	uss-se	with him /her/it	no
उन	un	that/those	no
उन्होंने	unhon-ne	they	no
उन्हें	unhein	them	no
उनके	unke	by them, their	no
उनकी	unkii	their	no
उनका	unkaa	their	no
उनसे	un-se	them	no
उनमें	un-mein	in them	no
यहाँ	yhaan	here	no
वहां	vahaan	there	no
यहीं	yaheen	here	no
वहीं	vaheen	there	no
ऐसा	eissa	like this	yes
वैसा	vaissa	like that	no
ऐसी	eissii	like this	yes
वैसी	vaisii	lke that	no
ऐसे	eisse	like this	yes
वैसे	vaise	like that	no
इन	inn	this	yes
इनके	inke	about them	no
इनमें	inmein	in them	no
यही	yahii	this/it	no
वही	vahii	that	no

Table 1. Demonstrative Pronouns and its indirect anaphoricity

S.No.	Case	Pronoun Forms	Pronoun Hindi
1	Nominative Case	iss	इस
2	Ergative Case	iss-ne	इसने
3	Accusative Case	iss-ko	इसको
4	Instrumental Case	iss-se, isse iss-ke	इससे, इसे, इसके
5	Dative Case	is-ko, isse	इसको , इसे
6	Ablative Case	iss	इस
7	Genative Case	iss-ka, iss-ki, iss-ke	इसका, इसकी, इसके
8	Locative Case	iss-mein, iss-par	इसमें, इस पर

Table 2. Case marking of pronoun "iss"

(2) इस प्रकार उक्त निर्देश के आलोक में दोनों आरोपियों ने आज अदालत के समक्ष आत्मसमर्पण किया तथा ज़मानत याचिका दायर की थी .
Iss prakaar ukt nirdesh ke alok mein dono aaropion ne aaj adalat ke samaksh aatmsamarpan kiya tataa jamaanat yachikaa daayar kii thii.
Thus, in the light of the above directions both accused surrendered to the court today and filed bail petition.

The presence of words like "prakaar", "tarah", "baabat", after "iss" intuitively conveys that the pronoun is indirectly anaphoric and will not have a referent in the discourse. Further the presence or absence of case form or connective also helps us in assigning the indirect feature to our demonstrative pronoun as shown in sentence (3).

(3) इसी सिलसिले में पुलिस को दो महिलाओं की भी तलाश है
issii silsile mein police ko do mahilaon kii bhii talaash hei.
In this context police is in search of two ladies as well.

The presence of "mein" (in) after "silsile" (context) also conveys that the demonstrative pronoun "issii" (this) is a modifier and is adjunct to the sub sentence "police is in search of two ladies as well". The pattern "prakaar" if followed by auxiliary verb "hei (be) is directly referential. Therefore the role of connectives becomes important in the definition of referentiallity. Two cases in our text appeared in this form as shown in sentence (4).

(4) संहिता की प्रमुख विशेषताएं इस प्रकार हैं-
Sahinta kii pramukh visheshtayen iss prakaar hein.

Key features of Code are as follows:

Pronoun in a modifier can also have a direct referent in the discourse as shown in sentence (5).

(5) इस संस्थान के कार्यालय में नये छात्रों के स्वागतार्थ एक समारोह का आयोजन किया गया।
Iss sansthaan ke kaaryalya mein naye chaatron ke swaagatarth ek samaaroh kaa aayojan kiya gaya.
In the honour of new students a function was organized in the office of this institution.

The presence of noun "sansthaan" (institution) after "iss" is indicative of direct anaphoric feature of "iss".

Our approach is based on the occurrence of certain collocation patterns. We look at the collocation patterns occurring after demonstrative pronouns, if they do not have a nominal which may have appeared earlier, we see if it can be inferred as indirect anaphor by searching for occurrence of certain patterns. Some of commonly occurring patterns are "iss prakaar", "iss tarah", "eissii baat" etc. These patterns refer to a semantic category. Based on different information structures the pronouns are classified in different semantic categories and thus provide addition information that for these pronouns search for the antecedent should not be performed. Zaidan et al. (2007) also advocated the use of such additional information in the corpus.

We hypothesize that cognitive status of patterns following the demonstrative pronouns or personal pronouns account for the difference in the anaphoricity of the pronoun. Such patterns are known as collocation patterns. Common usage of collocation patterns along with pronouns and identifying their relationship, support "natural" choices of referent. Prasaad et al. (2004) used role of connectives in the development of Penn Discourse Tree Bank (PDTB) and (de Eugenio et al., 1997; Moser and Moore, 1995; Williams and Reiter, 2003) in Natural language generation. The findings reveal novel patterns regarding the collocation patterns for discourse and suggest additional experiments.

3. Methodology

The process of semantic classification of indirect anaphora required (a) selection of a corpus in Hindi, (b) identification of features that differentiate direct anaphora from the indirect one, (c) validation of our proposal using machine learning approach, and (d) development of automatic classification system for indirect anaphora. Our corpus should be encoded using Unicode. Hindi text using fonts which we may not be able to process seamlessly across different platform are not preferred. Identification of specific features requires careful analysis of corpus and formulation of appropriate rules. Since the data set is small, validation of scheme requires a selection of suitable algo-

rithms. In this paper we shall address first three issues. Development of automatic classification system will be carried out after fine tuning of our annotation scheme.

3.1. Corpus selection

We consider the data from Emille corpus. The corpus is based on the news items from Ranchi express (Sinha, 2002) and is the only known corpus in Hindi. The study aimed at improving the corpus with the semantic annotation for indirect anaphora. We analyzed 177 news items having 1334 sentences, 1600 demonstrative pronouns of which 97 (12.44 %) were indirectly anaphoric. The corpus is annotated for anaphora using scheme based on (Botley and McEnery, 2001) and customized for Hindi. Further Botley (2006) has also pointed out the limitation of his scheme and urged to encode more information essential for understanding indirect anaphora. This motivated us to further look into the annotation scheme adopted for the corpus.

Each occurrence of demonstrative pronoun is coded in an XML-compatible format so that it could be extracted automatically from the text. The indirect anaphora in this corpus is annotated as inferable antecedent. These are the cases that can be derived from the discourse but explicit noun phrase does not appear in the text. However existing encoding does not allows us to apply the resolution algorithms, as the exact antecedent cannot be extracted from the corpus. Further the pronoun marked as a direct or indirect, does not specifies what actually distinguishes direct anaphor from the indirect ones. We propose an extended scheme for annotating the corpus on indirect anaphora and incorporate features, which help us in identifying the indirect anaphoricity behavior of the pronoun. For our study we have considered only those pronouns, which have been marked as Inferable. The Emille corpus is based on the news items from Ranchi express and is the only known corpus in Hindi annotated for anaphora. The corpus is annotated for anaphora using scheme based on (Botley and McEnery, 2001) and customized for Hindi corpus by (Sinha, 2002). Each occurrence of demonstrative pronoun is coded in an XML-compatible format so that it could be extracted automatically from the text. The indirect anaphora in this corpus is annotated as inferable antecedent. These are the cases that can be derived from the discourse but explicit noun phrase does not appear in the text as a referent. The existing encoding does not allows us to apply the resolution algorithms, as the exact antecedent cannot be extracted from the corpus. Further, the pronoun marked as a direct or indirect, does not specifies what actually distinguishes direct anaphor from the indirect ones.

We propose an extended scheme for annotating the corpus on indirect anaphora and incorporate features, which help us in identifying the indirect anaphoricity behavior of the pronoun. For our study, we have considered only those pronouns, which have been marked as Inferable. The choice inspired by the work of Brown-Schmidt et al. (2005); Eckert and Strube (2000), these features captures preferences for NP- or non-NP-antecedents by considering a pronoun's predicative context. The underlying

assumption is that if certain pattern occurs after personal or demonstrative pronoun, then the pronoun will be likely to have a non-NP-antecedent.

3.2. Corpus annotation scheme

Theories proposed (Gundel et al., 2005) presents the case of indirect anaphora in English texts as a case of focus and attention. Kerstin and S.Hansen-Schirra (2003) have presented the scheme of annotating indirect anaphora. All these schemes were presented for English where it, that and this are generally used for demonstrative pronouns and also behaves as an indirect anaphora. (Dipper and Zinsmeister, 2009) annotated German corpus based on the semantic restriction and contextual features derived from the corpus. Navarretta and Olsen (2008) developed annotated Danish and Italian corpus for abstract anaphora.

Since indirect anaphora is based on cognitive kinds of relations, the classification may not be agreed upon between different annotators. However to start with we describe our own classification based on collocation pattern preference reflecting the key specific feature of our text corpus. The generalized classification proposed in (Fan et al., 2005) is based on abstraction, name-entity-relation, attribute relation and associative relation. However for Hindi corpus we adopt the classification scheme guided by the collocation pattern and the case marking that follows. The rationale of using this scheme is to keep the annotation process simple yet useful. As long as the annotator is spending the time to study example and classify it, it may not require much extra effort for classification.

The annotation scheme deals with the manual annotation of pronouns without an explicit noun phrase antecedent. Direct anaphors are able to find antecedent from noun phrases, the indirect anaphors are classified based on the semantic relations. The semantic classification ranges from explicit relations derivable from the information present in the discourse to implicit relations based on pure inference.

We focus once again on demonstrative pronouns and the ones marked as inferable in the corpus. We look at the collocation patterns for pronouns. The most popular approach for locating collocation patterns is the window-based which collects word co-occurrence statistics within the, context windows of an observing headword to identify word combinations with significant statistics-as collocations. For our experiment we have used the Heidelberg Tenka text concordance tool, an open source text analysis software and extracted the collocation patterns along with the pronouns as a head word and annotated the text as shown in Table 1. If the pronoun is indirectly inferable than pattern following the pronoun is also encoded and the semantic type is also specified according to the semantic classification given in Table 3. An example of annotation is shown in Example 6.

Feature	Value1	Value2	Value3	Value4	Value5
Distance Marking	P (proximal)	D (Distal)	None	None	None
Nature of deixis	P (Pronoun)	D (Demonstrative)	Z (Zero)	None	None
Recoverability of Antecedent	D (Directly Recoverable)	I (Indirectly Recoverable)	N (Non-recoverable)	0 (not applicable, e.g.) exophora)	None
Direction of reference	A (anaphoric)	C (cataphoric)	0 (not applicable, exophoric or deictic)	None	None
Phoric Type	R (Referential)	0 Not Applicable	None	None	None
Syntactic Function	M (Noun Modifier)	H (Noun Head)	0 (Not Applicable)	None	None
Antecedent Type	N (nominal)	P (propositional/ Factual)	C (Clausal)	J (Adjectival)	O (None)
Pronoun pattern	Pronoun and subsequent construct in the sentence				
Case marker/ Connective	Case marking or connective following the pronoun				
Semantic/ category	semantic categories as defined in Table 5				

Table 3. Feature Set used for annotation

Patterns following pronouns
samjhaa, aarakshan, liye, prakaar, baat, dishaa, sthiti, jaankaari, tarah, ek, paristhiti, roop, tak, kram, dhandhe, kuch, paksh, alaava, sandarbh, arth, or , gambhirta, siidhaa, tatvon, silsile, silsila, prashikshan, sambandh, gambhiirta, dushparinaam, kadam, galat, badii, dushparinam, ghatna, kaaranon, tamam, baavjood, saath , tayaari, matlab, manzar, moukaa, katthinaaii, baabat, sarvoch, saare_aaropon, suvidha, hii, baare, vyavasthaa, maukaa, maamla, sandesh, charchaa, aalok, suvidhaa, kitnii, prashnon, sambadh, sanchaalan, aashye, saath-saath, maansikta, durust, hinsak, gervajib, naaraz, koi, nai, vistrit, maamle, charchaaen, laabh, saari, saare, kaarnon , vishleshnon, seet, kuchh, khade, tahat, anapekshit, asar, ghatana, mudde, par, bhayaaveh, to, train, tayaarii, sab, siidha, tamaam, kathinaaion, baavzood, null
Case marker and connectives
mein, par, ki, kii, ke, se, hii, ka, ko, null, O
Semantic Categories
event, act, object, emphasize, subset, result, adjective, equivalence, type, summarize, reason, situation, context, additional, information, undefined

Table 4. Annotation feature set used for semantic annotation

(6) <s tag=2>झारखंड सरकार ने लातेहार, सिमडेगा, सरायेकेला और जामताड़ा को आज जिला बनाने संबंधी अधिसूचना जारी कर दी । </s><s tag =3> < w c= 1, tag="P,D,In,A, R,M,O, iss, prakaar, null, summarize"> इस </w> प्रकार अब झारखंड में जिलों की संख्या १८ से बढकर २२ हो गयी है । </s>
<s tag=18> राज्य में नए प्रशासनिक इकाईयों के गठन के सम्बन्ध में निर्णय लेने वाली उच्च स्तरीय समिति ने बैठक करके चार नये जिले बनाने की सिफारिश भी की थी । </s> <s tag=19> राज्य के मुख्य सचिव वी. एम. दुबे <w c=6, tag="P,D,D,A,R,M,N,iss, _, _,_"> इस </w> समिति के प्रमुख हैं । </s>

3.3. Classification

In most of the cases where pronoun is indirectly referenced the pattern following the pronoun is normally an abstract form of noun phrase, or characterization of the information conveyed in the discourse. This characterization cannot be capturing through the explicit referent, but a semantic annotation does provide the information about the status of information so far present in the discourse. A partial list of patterns and possible classification used in our experiment is listed in Table 4. In most of the cases "prakaar" is classified as "summarization" but if "prakaar" is followed by "ka/ki" then it is classified as "equivalence". Also in some cases two different annotators may classify same pattern differently. "iss-ke saath hii" (along with this only)

could be classified as an "event" and an "emphasize" as well. For our present study we include both the cases in our experiment.

Let
- S: list of tokens of semantic classification
- C: list of case markers and connectives {hii, ka, kii, ki, se, mein, par,…}
- T: list of tokens { "prakaar, "tarah", "kram",…}
- D: list of pronouns directly inferable but not indirectly inferable {issne, ussne, ussko, issko,…}
- R: list of remaining pronouns (these pronouns exhibit both type of behaviour) {yeh, iss, uss, inn,…}
- L: $D \cup R$
- SI: classification $SI \in S$
- XL: list of pronouns in the corpus
- X: current pronoun from the list XL; $X \in XL$
- XP: pattern following X
- XC: case marking
- ST: string consisting of X, XP, XC
- SN: syntactic category
- N: noun
- P: pronoun

For given pronoun X
1. Through concordance obtain string S which includes X, XP and XC
2. If $X \in D$ then skip to the next pronoun (pronouns defined purely for direct anaphora are eliminated from our study)
3. If a pronoun X is of noun type N and if the collocation pattern $XP \in T$ is an elaboration of one of the form from the classification list S then go to step 4
4. If a pronoun X is a modifier and if the collocation pattern XP following the pronoun X is an elaboration from one of the elements in classification list S, the pronoun is indirectly inferable.
5. If step 2 or step 3 is true then look for the connective/case marker $XC \in C$. If condition is satisfied annotate the given pronoun with X, XP, XC, SI along with other annotation provided in the Emille corpus else keep these features "null".

Classification rules
Since our classification scheme is based on the semantic cues provided by the concordance patterns of a discourse segment whose head is the pronoun with non NP-antecedent, we exploit this information for the purpose of classification. We have framed 25 rules, which can be applicable to a specific pronoun in a discourse. Some of the rules are given below:

Rule 1
IF : SN in H $\wedge$ PRONOUN in{iss} $\wedge$ XP in {prakaar} $\wedge$ XC in {null} $\Rightarrow$ CLASS = result

Rule 2
IF : SN in M $\wedge$ PRONOUN in {issii} $\wedge$ XP in {prakaar} $\wedge$ XC in {ka} $\Rightarrow$ CLASS = type

Rule 3
IF : SN in H $\wedge$ PRONOUN in {iss, issi} $\wedge$ XP in {tarah} $\wedge$ XC in {ke, ka} $\Rightarrow$ CLASS = type

Rule 4
IF : SN in M $\wedge$ PRONOUN in {iss, eisse} $\wedge$ XP in {tarah, tatvon, tamaam} $\wedge$ XC in {ki, kii, ke, ka, null} $\Rightarrow$ CLASS = type

Rule 5
IF : SN in M $\wedge$ PRONOUN in {ussii} $\wedge$ XP in {roop} $\wedge$ XC in {mein} $\Rightarrow$ CLASS = type

Rule 6
IF : SN in M, H $\wedge$ PRONOUN in {issii} $\wedge$ XP in {tarah} $\wedge$ XC in {null} $\Rightarrow$ CLASS = equivalence

Rule 7
IF : SN in M $\wedge$ PRONOUN in {issii, inn} $\wedge$ XP in {prakaar, saare} $\wedge$ XC in {se, null} $\Rightarrow$ CLASS = equivalence

Rule 8
IF : SN in M $\wedge$ PRONOUN in {ussii} $\wedge$ XP in {tayaarii} $\wedge$ XC in {ke} $\Rightarrow$ CLASS = adjective

Rule 9
IF : SN in M $\wedge$ PRONOUN in {inheen} $\wedge$ XP in {kaarnon} $\wedge$ XC in {se} $\Rightarrow$ CLASS = reason

Rule 10
IF : SN in M $\wedge$ PRONOUN in {issii} $\wedge$ XP in {paksh} $\wedge$ XC in {ki} $\Rightarrow$ CLASS = subset

Rule 11
IF : SN in M, H $\wedge$ PRONOUN in {yeh, iss, issii} $\wedge$ XP in {ek} $\wedge$ XC in {mein, ka, nom, null} $\Rightarrow$ CLASS = emphasize

Rule 12
IF : SN in M, H $\wedge$ PRONOUN in {yeh, iss, isse, issii, iss-ke, eisaa, eisse} $\wedge$ XP in {kram, gambhirta, silsile, silsila, ghatna, manzar, maamla, kuchh} $\wedge$ XC in {mein, ke, hii, ka, null} $\Rightarrow$ CLASS = event

Rule 13
IF : SN in M, H $\wedge$ PRONOUN in {iss, isse, isskii} $\wedge$ XP in {samjhaa, jaankaari, sambandh, baare, ghatana} $\wedge$ XC in {mein, kii, null} $\Rightarrow$ CLASS = information

When the pronoun has a direct NP-antecedent in the discourse the classification is categorized as direct only and pattern feature and case marker feature are not analyzed. The classification obtained suggests that the use of dictionary and thesaurus would improve the classification scheme.

Few examples of classifications based on the above rules are listed in Table 5.

Classification	Example
Event	जंगल बचाने का अभियान यहीं तक जारी नहीं रहा
Act	इस दिशा में चलाया जा रहा कार्य
Emphasize	यह एक सोची-समझी
People	इसी पक्ष की जांच-पड़ताल
Result	इसके लिए हमें मिलजुल कर कार्य करना होगा
Adjective	उसी तैयारी के साथ
Equivalence	इसी तरह की अन्य जातियां भी हैं
Type	इसी प्रकार का अधिकार
Summarize	इस प्रकार अब झारखण्ड में
Reason	इन्हीं कारणों से
Situation	ऐसी स्थिति का विरोध किया
Context	इन सन्दर्भ में
Additional	इसके बावजूद दूः स्थिति है कि
Information	इसकी जानकारी नहीं मिली

Table 5. Patterns and Classification for semantic annotation

3.4. Experiment

The distribution of anaphors with NP-antecedent (12.44 %) and non NP-antecedents (12.44 %) in Emille corpus is shown in Table 6. This figure is comparable to the number of pronouns without NP antecedents as reported in Gundel et al. (2005) as 16 % for New York times corpus, Poesio and Viera (1998) as 15 % or their corpus and Botley (2006) as 20 % for Associate Press corpus. All these studies are for English texts. We understand that this feature is similar across languages.

Though the present work deals with developing semantic annotation scheme for indirect anaphora in Hindi, the corpus obtained can be used for developing automatic classification models. (de Eugenio et al., 1997) has also applied the feature-based information in discourse for automatic generation of explanation in text generation. In our case the automatic classification of semantic categories can be used to automatically derive anaphora rules and ultimately use in anaphora resolution system. This will also prevent the human subjectivity, which is the main limiting factor in the de-

Pronouns	**direct**	**indirect**
yeh	184	11
iss	275	32
isse	23	2
issii	27	16
Iss-ka	18	1
isskii	15	1
issmein	12	1
usii	14	5
eisaa	29	2
eisee	13	11
eisse	23	4
yaheen	1	1
inn	47	1
inheen	2	1
Total	683	97
780	87.56 %	12.44 %
Total sentences: 1334		
Total demonstratives: 1600		

Table 6. Distribution of pronouns

velopment of large and reliable corpus. Two annotators may have different views about the category to which the given utterance should belong (Reiter and Sripada, 2002). We also experienced these problems in our attempts to tag the Emille corpus, which initially had some bugs, and our annotation was also based on our judgement, which cannot guaranty same results all time. This complexity of anaphor classification made us experiment with machine learning approaches.

After having tagged the data set it was easier for us to experiment with these methods. After trying several algorithms we chose to experiment with JRIP, J48 (the Weka implementation of C4.5) and LMT (Logical Model Tree)(Witten and Frank, 2005). First experiment included all the occurrences of demonstrative pronoun (with NP-antecedent and non NP-antecedents). Performance of J48 a C.45 decision tree based algorithm at confidence factor 0.8 improves to 88.462. Algorithm J48 computation time is far less than the LMT algorithm. Where J48 builds model in 0.02 seconds LMT algorithm 147.47 seconds. This makes J48 a preferred algorithm for very large

datasets. But since our corpus size is small, LMT gives a better model as it combines the advantage of regression and tree approach.

Data Set	JRIP			J48			LMT		
	S(%)	K	E	S(%)	K	E	S(%)	K	E
100	83	0.4684	0.0271	85	0.6488	0.147	84	0.6277	0.0310
200	86.57	0.6205	0.0227	88	0.7182	0.012	88	0.7213	0.0131
300	81.4545	0.4925	0.0293	86.5455	0.7073	0.0148	86.5455	0.7075	0.0978
400	82	0.4376	0.0277	86.5	0.6715	0.0143	85.75	0.6571	0.0155
500	85.7692	0.4202	0.0219	88.462	0.6598	0.0113	89.2308	0.6732	0.0116

E-Mean absolute error
S-Success Rate
K- Kappa Statistic

Table 7. Performance Measures of algorithms on given data sets

4. Analysis

The analysis of the experiment suggests that the performance measure in the current data set is dominated by the directly inferred pronouns. Experiment with the dataset excluding directly inferable pronouns resulted in a considerable drop in the performance in LMT from 89 % to 55 %. Performance of JRIP and J48 falls to 39 % and 42 % respectively. For reliable results, getting sufficiently large corpus is difficult. Further the linguistic cues used for the semantic classification of indirect anaphora needs further investigations as patterns like “prakaar”(10.31 %) and “tarha” (11.34 %) account for the major contribution toward the indirect referentiality of pronoun but other patterns like “tatvon”, “sthiti” and many others had marginal number of instances. Some patterns appeared only once. Other factor that we have ignored is the presence of words from other languages like English, which is becoming the natural way of communication and thus making the task of text processing more difficult.

The other solution could be the refinement of rules with usage of thesaurus in deciding the semantic classification, associating weight factor to positive classification and penalties for incorrect classification and specifying met rules. Further two annotators may also differ in their judgment about the class association. This would result in two different corpora for the same text. Also the annotator himself may not be able to decide exact category. In such cases either we may allow multi membership or assign different weights to the assignment. The possibility of inclusion of the indirect pronoun in different categories results in conflict in the present scheme. This conflict can be improved by incorporating a score value to each classification as follow: Premise of the rule $\Rightarrow$ { Class, likelihood} Where likelihood takes values as in the

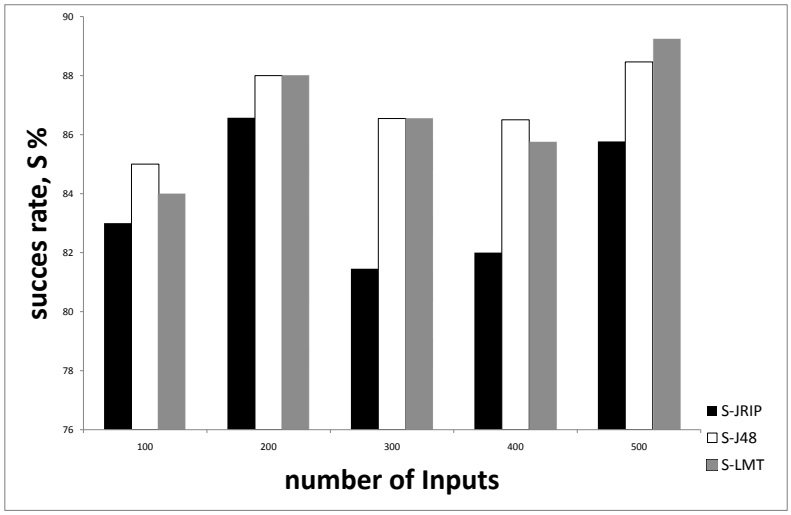

Figure 1. Success Rate of Algorithms on varied size of data sets

range of { -10 to +10 } ; positive value is for the likelihood of the correct classification, whereas negative values are indicative of the penalty of wrong classification.

Expanded rule specification could be Premise of the rule $\Rightarrow$ { ($Class_1$, $likelihood_1$), ($Class_2$, $likelihood_3$),..., ($Class_n$, $likelihood_n$) }.

Expanded rule can include the likelihood of class association for all classes. This requires more detail study of the corpus to decide upon exact likelihood values. In the present corpus the amount of instances available for indirect anaphora is too less to conclude concretely from the results obtained. Another possible solution is reduction in the number of classes by merging some of the categories. But in that case the extraction of semantic, which is useful in text cohesion, will be lost.

5. Conclusion

In this paper we have presented an enhanced annotation scheme on Emille corpus for indirect anaphora in Hindi. Annotation is enhanced with the semantic information for indirect anaphora. We experimented with automated classification using machine-learning approaches and our results show that the semantically enhanced annotation is a rich source of information for natural language understanding and

generation systems and for conducting data oriented research. Though the present model does not produce desirable results, fine-tuning of rules, incorporation more rules and with more data set better performance can be achieved.

Bibliography

Botley, S. and A. McEnery. Demonstratives in English: a corpus-based study. *Journal of English Linguistics*, 29:7–33, March 2001.

Botley, S. P. Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1):73–112, 2006.

Boyad, A., W. Geeg-Harison, and D. Byron. Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In *ACL Workshop on Feature Engineering for Machine Learning in NLP*, pages 40–47, Ann Arbor, June 2005. Association for Computational Linguistics.

Brown-Schmidt, S., D.K. Byron, and M.K. Tanenhaus. Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language 53 (2), pp. 292–313*, pages 292–313, 2005.

de Eugenio, B., J.D. Moore, , and M. Paolucci. Learning Features that Predict Cue Usage. In *ACL/EACL 97*, 1997.

Dipper, S. and H. Zinsmeister. Annotating Discourse Anaphora. In *Third Linguistic Annotation Workshop*, pages 166–169, Suntec, Singapore, August 2009. ACL-IJCNLP.

Eckert, M. and M. Strube. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics 17 (1)*, pages 51–89, 2000.

Fan, J., K. Barker, and B. Porter. Indirect Anaphora Resolution as Semantic Path Search. *KCAP' 05*, October 2005.

Gasperin, C. and R. Viera. Using word similarity lists for resolving indirect anaphora. In *ACL Workshop on Reference Resolution and its Applications*, pages 40–46, Barcelona : Copisteria Miracle, S.A., 2004.

Gelbukh, A. and G. Sidorov. Word choice problem and anaphora resolution. *ISMT-CLIP*, 1999.

Gundel, J., N. Hedberg, and R. Zacharski. Pronouns without NP Antecedents: How do we know when a pronoun is referential. *Anaphora Processing: Linguistic, Cognitive and Computational Modelling, ed. by Antonio Branco, Tony McEnery, and Ruslan Mitkov. John Benjamins*, pages 351–364, 2005.

Gundel, J., N. Hedberg, and R. Zacharski. Directly and Indirectly Anaphoric Demonstrative and Personal Pronouns in Newspaper Articles. In *Proceedings of the Sixth Annual Discourse Anaphora and Anaphora Resolution Colloquium*, 2007.

Kerstin, K. and S.Hansen-Schirra. Coreference annotation of the tiger treebank. In *Workshop Treebanks and Linguistic Theories 200*, pages 221–224, 2003.

Mitkov, R. Factors in Anaphora Resolution: They are not the Only Things that Matter. A Case Study Based on Two Different Approaches. In *Proc. of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 1997.

Mitkov, R. *Anaphora Resolution*. Longman, London, 2002.

Moser, M.G. and J. Moore. Investigating Cue Selection and Placement in Tutorial Discourse. In *ACL95*, 1995.

Navarretta, C. and S. Olsen. Annotating abstract pronominal anaphora in the DAD project. In *REC-2008*, May 2008.

Pandharipande, R. and Y. Kachru. Relational Grammar, Ergativity, and Hindi-Urdu. *Lingua*, 41:217–238, 1977.

Poesio, M. and R. Viera. A corpus-based investigation of definite description use. *Computational Linguistics*, pages 183–216, 1998.

Prasaad, R., E. Miltaski, A. Joshi, and B. Webber. Annotation and Data Mining of the Penn Discourse TreeBank. In *ACL Workshop on Discourse Annotation*, July 2004.

Reiter, E. and S. Sripada. Human Variation and Lexical Choice. *Computational Linguistics*, 28 (4):545–553, 2002. ISSN 0891-2017.

Schwarz, M. Establishing Coherence in Text. Conceptual Continuity and Text-world Models. *Logos and Language*, 2(1):15–24, 2001.

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

Williams, S. and E. Reiter. A Corpus Analysis of Discourse Relations for Natural Language Generation. In *Corpus Linguistics*, 2003.

Witten, I. H. and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.

Zaidan, O., E. Jason, and C. Piatko. Using annotator rationales to improve machine learning for text categorization. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 260–267, Rochester, NY, April 2007.

Address for correspondence:
Kamlesh Dutta
kd@nitham.ac.in
National Institute of Technology
Hamirpur (HP)-177005, INDIA