Anaphora Resolution in Hindi

A Project Report submitted in fulfillment of the requirement for the

award of the degree of

Master of Technology

in

Computer Science & Engineering

Under the Supervision of

Dr. Rajni Mohana
(Supervisor)
By
Ashima

Enrollment No: 142204

Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh-India 173234

Certificate

This is to certify that the work contained in thesis “Anaphora resolution in Hindi”, submitted by Ashima (142204) in partial fulfillment for the award of degree of Master of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been made under my supervision.

This thesis has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date:

Dr. Rajni Mohana (M.Tech Supervisor)

Assistant professor (Senior Grade)

Acknowledgement

I would like to thank all the persons in my life for their support and encouragement for my research. First of all, I would like to thank my advisor Dr. Rajni Mohana for constant guidance, encouragement and support. Without guidance and motivation from her, my research would not have been possible. Her deep knowledge of linguistics, NLP or other subjects has always helped me, not only towards my research but also in other domains of life. Her dedication towards her subjects and work has inspired me to remain dedicated towards the goals for life. I would also thank to my family members, father Mr. Ashwani kukkar, mother Mrs. Ranjna, nana g Sh. Rajinder Pal Soni and younger brother Kushal who always supported me in every way possible. I would like to thank JUIT members H.O.D Dr S.P ghrera, Co-ordinator Dr. Pradeep kumar and Ph.D scholar Ms. Sukhnandan Kaur for continuous support. Next I would like to thank all my seniors and juniors. I am specifically thankful to Abhilasha, Ishani, Nitish and Ravi for all the continuous support, suggestions and criticism during my stay in JUIT. Time spent with you people has changed my thinking towards life and world. Life without you guys would have been boring and tedious throughout these years. Finally, I would also like to thank all my classmates who have made this journey easier.s

Date: ______________

Signature: ______________

Ashima (142204)

Table of Content

Sr. No Topic Page No.

Chapter:1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

List of Tables

ABSTRACT

Natural language processing (NLP) is way of understanding the human languages by machine which involves the field of computer science and artificial intelligence. The basic aim of NLP is to develop such software that will analyze, understand and generate human languages. It has many challenges like word sense disambiguation, sentiment analysis, summarize from paragraph to sentences, correct the search query, language detection, missing apostrophes, etc, anaphora is one of them. Anaphora is considered as the problem of referring pronoun phrase to noun phrase with n the sentence or between the sentences. Anaphora resolution is required in various NLP applications like auto summarization system, machine translation system, question answering system etc. Lot of work has been done in English but less work has been done in Hindi. Anaphora Resolution in Hindi is a complex task because Hindi is free word order language. It is based on the verb so pronouns in Hindi show a great deal of dubiety. This thesis presents the various approaches like Ruble based, Machine based, Learning based etc to resolve various issues of anaphora resolution in Hindi. Firstly, the thesis proposes an algorithm by using rule based approach to resolve pronouns based on gender and number agreement which gave 79% of accuracy in terms of F-Score.

Secondly, the hybrid based approach which is a combination of Rule based and Learning based with dependency structures for anaphora resolution in Hindi to resolve co reference, number and gender agreement, animacy issues. The performance of our system is checked by using MUC, B3, CEAF, F- score on the different data set and overall performance is 84.50%, 78.9%, 78.03% and 78.06% respectively, exhibits significantly different behaviors .In the approach, a rule-based module is used to resolve simple anaphora, while learning tools are used to resolve more ambiguous anaphora using grammatical and semantic features. The results show that, use of dependency structures provides syntactic knowledge which helps to resolve some specific types of references. The experiments show that even with limited data and using rules and training for Hindi language, reasonable resolution accuracy can be achieved for anaphora resolution

CHAPTER 1: INTRODUCTION

NLP is the natural language processing is referred with the intercommunication between computers and human languages [1]. It is associated with area of artificial intelligence and computer science. The basic aim of NLP is to develop such software which can understand various human languages as well as analyze and generate the languages that human used naturally so that a machine can interact with the human by using simple human languages. This target is difficult to achieve as it requires to grasp the aware about the concepts of words or phrase (what they stands for) and how these concepts are linked together in a meaningful way. It is easy for a human to understand, learn and use the human languages, symbol system but it is hard for a computer to master it. Now-a-days machines have the ability of upturning the large matrices with speed and grace but they still fail to master the basics of our written and spoken languages to the machine.

The representation of the semantics of the text in computing is very important component. Initially we need to define the notation of the representation else it may lead to ambiguity. Syntactic structure describes how the words in the sentence are related to each other. Its representations of the languages are based on the notion of context-free grammars. Syntactic representation represents sentence structure in terms of phrases which are the subparts of other phrases usually in a tree form. These structures give the details on the parts of speech for each word and phrases. The Logical form defines the representation of the context-independent meaning of a sentence. It encodes the all possible word senses then identifies the semantic relationships between words and phrases. Natural language system uses the general knowledge for representing and reasoning. It has many challenges in representation like word sense disambiguation, sentiment analysis, summarize from paragraph to sentences, correct the search query, language detection, missing apostrophes etc and Anaphora is one of them because it is difficult to address your computer as though you were addressing another person[2]**.**

Anaphora is defined as the problem of pointing the pronoun phrase to the noun phrase within the sentences or between the sentences. In linguistics, anaphora is defined as the use of an expression which refers to another expression in its context. The ‘pointing back’ word or phrase is called anaphor and the entity to which an anaphor refers or for which it stands is its .

E.g. : ***Sachin asked Rahul to borrow him the money.***

- The referring word (him) is called an anaphor.
- The preceding form (Sachin) is called an antecedent.
- The document in which the anaphor fall is called its discourse.
- Anaphora resolution is the procedure of directing the antecedent of an anaphora.

Now in the above example ‘Sachin ‘, ‘Rahul’ are the noun and Him is the pronoun. ‘Him’ pronoun is referring to the entity ‘Sachin’. The process of identification of the referent is known as ‘Anaphora Resolution’.

While there has been remarkable research for anaphora resolution in English and other foreigners languages but a limited amount of work is done for Indian languages. In this thesis, we aim to resolve the various issues of anaphora resolution in Hindi and explore linguistic features which are helpful in the resolution process.

Anaphora resolution can be categorized into Intersentential and Intrasentential [3] :-

- **Intrasentential** is the expression where the antecedent and its anaphor fall in the same sentence .
- **Intersentential** is the expression where antecedents

and its anaphor fall in different sentences ..

It can also be classified as [4] :-

- **Entity** anaphora defines those pronominal references which refer to a real Entity such as Person, place and other common nouns.
- **Event** anaphora refers to those pronominal references which refer to Events.

Now challenge in anaphora resolution is

Challenge: *Who does each pronoun refer to?*

e.g.:

S1: The books were given to the students because they were required.

S2: The books were given to the students because they were useful.

S3: The books were given to the students because they were there.

Who does each *they* refer to in the example? Because of the different interpretation of "*they*" in each sentence, the wholesome meaning of the sentence changes.[5]

In S1 – '*They*' refer to '*students*', S2- '*They*' refer to '*books'* and in S3 –'*They*' refer to '*students'*. A human can easily understand this but a machine can't.

Anaphora occurs very repeatedly in text and spoken sentences. The main aim of natural language processing applications is that they must resolve anaphors but there is no existing theory or methodology which resolves all anaphora. Almost all the NLP applications require anaphora resolution.

1.1 Applications of NLP

There are many applications of the natural language processing which require resolving the pronoun resolution that are:

- **Automatic summarization system:**

It is the process of produce the readable summary from the large text document which contain the most important points of the original text document[6]. Anaphora resolution plays a very important role in the summarization system. For example

e.g**. "** ***Anu who was born to Punjab , served as the president of India from 1993 to 1997*** **"**

This sentence requires the knowledge of the referent of pronoun 'She' [7]. The summarization with solving the anaphora is impossible.

- **Question answering model:** It is process of finding the answer of the human language question from the given document [8]. According to [9] Anaphora resolution is an important task for searching the answer in the Question answering model. For explain this take a example below:

" Anu Walia who was born to Punjab . She served as the president of India from 1993 to 1997 **"**

In given sentences there is a query as "Who served as President of India from 1993 to 1997?" Now to obtain the answer, the system must have the knowledge about the pronoun 'She' is pointing to another entity in the previous sentence i.e. 'Anu Walia'.

- **Natural language generation system:**

It is the process of converting the information from computer databases into readable human language. It also required the knowledge of the referent of the pronouns to nouns.

- **Machine translation system:**

It is the process of translating the text from one language to other with the help of machines and software. Anaphora resolution is required to translate the language from one to another. Because in some languages the pronouns has the different forms according to their morphological properties (such as animacy, number, gender etc) of the words which require the knowledge of the referent of the pronouns to translate the pronoun form one language to another and it is come from the Anaphora resolution [9].

Information extraction system, name entity reorganization, discourse analysis etc., also require complete identification and complete plan to resolve an anaphora.

Anaphora resolution is not an easy process. The machine should have good understanding of the interaction between the syntax, semantics and pragmatics of a language. It requires not only knowledge but also expertise of almost all language processing domains as shown in (Figure 1.1).

1.2 Types of Knowledge

There are five types of knowledge which are requiring by the anaphora resolution to resolve the pronouns are described in Figure 1.1 and explanation is given below:

i. **Morphological and lexical knowledge** – It is process of the formation of the word. It includes the study of the structure of formation of words by the combination of sounds into minimal distinctive units of meaning that is called morphemes. Morphological knowledge is related to how the words are constructed from the morphemes.

ii. **Syntactic knowledge** – It is the process of combing the words to form the phrases, phrases combine to form clauses and clauses are combining to make the sentences. Syntactic analysis concerns with the formation of sentences by determining the words can be put together to form the correct sentences. It also determines the structural role of each word that plays in the sentence.

iii. **Semantic knowledge** - It is the process of determining the meanings of the words and sentences. Semantic knowledge is the study of context independent meaning of the sentences there is no matter in which context it is used. Due to the ambiguities the defining the meaning of a sentence is very difficult.

iv. **Discourse knowledge** – It is the process of determining the connected sentences which are bigger than a single sentence. It is concerned with the inter-sentential and intra-sentential links which are within the sentences or in different sentences.

v. **Real-world knowledge** - It is nothing but day to day knowledge that all speakers share about the world. It includes the general knowledge about the structure of the world and what each language user must know about the other user's beliefs and goals. This makes the language understating much better.

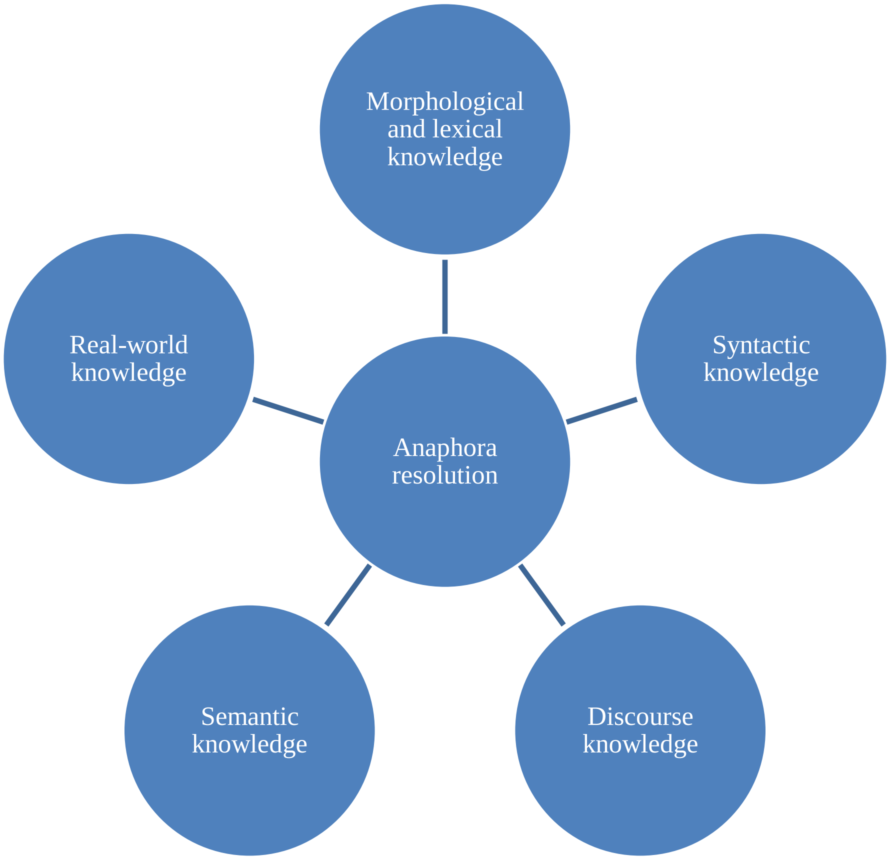


Figure 1.1: Different types of knowledge

The figure1.1 represents different types of knowledge's for resolving the anaphora and its antecedents. These knowledge's help in describing in structure of the sentences and the meaning of sentences and words

1.3 Classification of Anaphora and Pronouns in Hindi

Hindi is a part of the Indo-Aryan language family. In India there are 22 official Indian languages which are spoken among the 180 billion native speakers. Hindi is a free order word language which has the property of rich morphology. Hindi has no proper structure of the Subject- Object-Verb. It has default structure of SOV. Hindi is a verb final language that is called the ergative language. In Hindi language there is a agreement between the nominative NP and the verb. In this section we discuss the various pronouns that are used in Hindi like Possessive pronoun that can take place of noun phrase to show ownership and which has first, second and third person pronouns, Demonstrative pronoun that point to specific things, Reflexive pronouns that mention to particular subject of the sentence , Place pronoun that refer to location or places and Relative pronouns that introduce dependent clauses in sentences. Pronoun in hindi exhibits a great deal of confusion. Pronoun in the first, second, and third person do not convey any information about gender. In Hindi Language there is no difference between 'he' and 'she'. 'veh' is used for both the gender and is decided by the verb form. some forms, like 'usne'(he) 'usko'(him), are unambiguously singular but some forms can be both singular and plural, like 'unhone' (he)(honorific)/they, or 'unko'(him)(honorific)/ them. These all pronouns are very important for drafting the algorithms and rules for Anaphora Resolution [10]. The summary of comparison of pronominal anaphora for first, second and third person paradigm in Hindi is given below in table 1.1:

Table 1.1: Different types of pronoun in Hindi

S.r. no	Pronoun	Singular	Plural
1	**Possessive pronoun (Svatvātmāka sarvanāma)** A possessive pronoun is a pronoun that can take the place of		

	a noun phrase to show ownership		
	First Person (उत्तम पुस्ष)	मैं	-
		मझको, मुझे	-
		मेरा (m) , मेर (f)	मेरे (pl)
		-	हम
		-	हमको
		हमारा (m) ,हमार (f)	हमारे (pl)
	Second Person (मध्यम पुरूष):	तुम , आप (r)	-
		तुमको , आपको (r)	-
		तु)हारा(m) , तु)हार	तु)हारे (pl)

		(f) आपका (m) , आपकी (f)	आपके (pl)
	Third Person (अन्य पुरुष)	वह	वे(r)
		यह	ये (r)
		उसको, उनको(r)	उसको, उनको(r)
		इसको, इनको(r)	इसको, इनको(r)
		उसका (m) , उसकी(f)	उसके (pl)
			वे
			उनको
		उनका (m) , उसकी (f)	, उनके (pl)
2	**Demonstrative Pronoun (*Nishchyavaachak Sarvanaam*) :** Pronouns that point to specific things	यह	
		यह	
		वह	

		वह	
			ये
			वे
3	**Reflexive pronouns (Nijavaachak Sarvanaam):** Pronouns that refer back to the subject of the sentence or clause. They either end in –self	आप (or अपने आप), खुद, and स्वयं/स्वयम	खुद
4	**Relative pronouns (*Sambandhvaachak Sarvanaam*):** Pronouns that introduce dependent clauses in sentences	जब … तब/तो , जहाँ वहां , जैसे वैसे , जैसा वैसा, जितना उतना, जो वह	जब … तब/तो , जहाँ वहां , जैसे वैसे , जैसा वैसा, जितना उतना, जो वह
5	**Place pronouns** (जगह सवर्नाम) : Pronouns that refer to location	वहां , यहां	-

1.4 Different Approaches Used In Anaphora Resolution

Researchers have used various approaches to identify the various pronouns as shown in table 1 .Various models which are used are rule based, corpus based and applying AI and machine learning techniques [11]. The researchers have also combined these approaches to give better results in some cases. Those are called Hybrid approaches. These approaches can be broadly categorized into four types. They are shown under:

i. Rule Based Approaches

Rule based approaches combine knowledge sources and various factors that remove the items which are not required up to a set of the feasible items is obtained. The constraints work as a filter to remove the unwanted candidates within a set of defined rules. Thereafter, preference- based factors are applied.

Various Examples of algorithms under the Rule based approach:

Tree search algorithm (1978, Hobb) [12]

Shallow processing approach (1987,Carter) [13]

Multistrategy approach (1988, Carbonell and Brown) [14] [15]

Syntax based approach (1994, Lappin and Leass) [16]

Combination of linguistic and statistical methods (1996, Mitkov) [17]

ii. **Corpus Based Approaches**

Corpus approaches are those which use the available corpus (a collection of written texts). These corpuses have been created specifically for the discourse task.

Various Examples of algorithms under the Corpus based approach:

Knowledge-independent approach (Nasukawa 1994) [12]

Statistical/Corpus processing approach (Dagan and Itai (1990)[18].

Machine Learning Approach (Connolly, Burger and Day, 1994) [19]

iii. **Knowledge Poor Approaches**

Knowledge poor approaches are economical, fast and reliable. They are acceptable for huge category of languages because they do not depend on the semantic and syntax. They enroll many Artificial intelligence techniques for e.g. neural network system, semantic framework etc and do not depend on the domain and linguistic knowledge. They can be run without the parsing.

Various Examples of algorithms under the knowledge poor approach:

Kennedy and Boguraev, 1996 [20] [21]

Robust knowledge poor approach (Mitkov, 1996- 1998)[19][24]

COGNAIC (Baldwin, 1997) [25]

ROSANA (Stuckardt, 2001)[23][24]

iv. Discourse Based Approaches

Discourse is modeled through a sequence of sentences. A single item is focused at the any given point in the text and it has to be unlike from all others items that have been raised. In order to resolve anaphora, the world knowledge and conclusion are also employed.

Various Examples of algorithms under the discourse based approach:

Centering theory (Grosz et al, 1995) [26]

BFP algorithm (Brennan et al, 1987) [27]

S-List algorithm (Strube, 1998) [26]

LRC algorithm (Tetreault, 2001)[28]

1.5 Issues in Anaphora Resolution

There are various other salient factors i.e. issues which are considered by the researches while resolving anaphora. They are as under:

i. Recency factor :

It describes the pronoun in the current sentence referents to that noun which have foremost weights than those in the precursory sentence. [29] e.g.

"इशानी ने मोर देखा। वह बहुत सुंदर था।"

In the above example the pronoun "वह" can either refer to "मोर" or "इशानी".But according to Recency "मोर" is more close to "वह" as compare to noun "इशानी", therefore pronoun "वह" will refer to "मोर".

ii. **Animistic Knowledge:**

It filters the items which are based on the living beings. Inanimate items are discarded from lists which are non-living beings. Consider the following e.g.:

“कुशाल हर दिन खाना खाता था और अपनी पत्नी को भी खिलाता था| ”

In the above example pronoun “अपनी” mention to noun “कुशाल” as pronoun “अपनी” is an animistic pronoun. It always refers to animistic noun.

iii. **Gender Agreement:**

Gender Agreement differentiates the gender of the mortal with respect to the gender needed by the pronoun which is being resolved. Any item that does not meet the requirements is removed. e.g.:

“कुशाल ने दुकान से खिलौना खरीदा । वह उसे पसंद करता है|

इशानी ने मेले से खिलौना खरीदा । वह उसे पसंद करती है|”

In Hindi Language verbs are used to resolve pronouns based on gender agreement. In theAbove example from the verbs “करता है” and “करती है”, it can be understand that “उसे” refers to male and female respectively.

iv. **Number Agreement:**

The style of speech is checked for the singularity and plurality. Whether the item is plural but if the present pronoun which is solved doesn’t designate a plurality then the item is discard from list. The same process is used for singular items. e.g.:

“गुरु और चंदन अच्छे दोस्त हैं| और वे बहुत शरारती हैं|”

In the above example pronoun “वे” refers to “गुरु और चंदन” that is plural.

v. **IntrasententiaI and Intersentential sentences** [3] **:**

Intrasentential sentences: is the case in which the antecedent of the anaphor are in the same sentence e.g.:

"सीता कहती है कि वह कल तंजाउर जाएगी।"

Intersentential sentences : is the case in which the antecedents of the anaphor occurs in the a different sentences . In these sentences, e.g.

"राधा पंडित की बेटी है

उसे पूजा करना पसंद है।"

vi. <u>Pronoun resolution:</u>

Pronouns in Hindi shows a great deal of dubiety [6]. Pronouns in the first, second and third person do not show any information about the gender. There is no difference between "he" and "she" in Hindi language. e.g.:
"veh" is used for male / female and difference in gender is decided by verb form. "usne"(he), "usko" (him), are apparent singular but in few areas they can be both singular and plural, when we want to give respect we used "unhone" (he) /they, or "unko" (him) (honorific)/them.

vii. <u>Named entity recognition</u> :

It is the process in which Named Entities or nouns are recognized and then categorized into dissimilar classes of Named Entity classes. [30]
For example:

Organization, Name of Person, River, Sport, Country, State, city, etc.

e.g.:

"सीता/***person*** कहती है कि वह कल तंजाउर/***city*** जाएगी"

1.6 The basic process of Anaphora Resolution in Hindi

There are various processes to resolve the issues of anaphora resolution but the underlying methodology is same. The basic methodology is shown in (Figure1.2) – It composes of following sentences:-

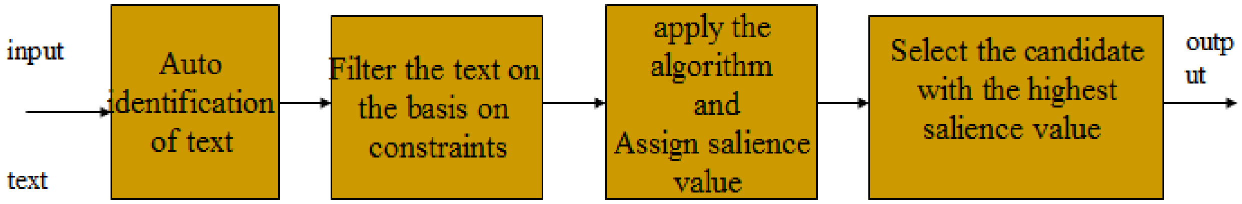

input
text
Auto identification of text
Filter the text on the basis on constraints
apply the algorithm and Assign salience value
Select the candidate with the highest salience value
outp ut

Figure1.2: The basic process of anaphora resolution

The figure 1.2 describes the basic process of resolving the pronoun by the machine. The working of the figure1.2 is explained below:

i. **Auto identification of text:**

The text is given as input and relevant discourse (paragraphs or sentences) is auto identified by machine where the plausible set of referents may exist.

ii. **Filtering of text:**

The constraints (rules or any other criteria) are applied to filter the text and the candidates are eliminated that are failed to meet the conditions.

iii. **Apply algorithm and salience value:**

The algorithm (approach) is applied to each of the remaining text or candidate and salience values are assigned to it .Select the candidate:- The candidate which has foremost salience value is selected as output.

1.7 Terminologies

In this thesis we have examined the anaphora resolution which involved various methods and terminologies in details in chapter 1, section 1.4, 1.5 and 1.6. In this section we have discussed the important and relevant terminologies which are used in this thesis.

1.7.1 Chunks

The breaking down of closely related words is called chunks [31]. It has not any syntactic structure. It is a fragment of the information that is used as input in the system. As we know Hindi is not depend on the order of the words and it is verb based language. So chucks play very important role in the dependency tree structure representation of the Hindi language. Chucks are the small units therefore the components of the chunk do not have any internal movement. The dependency structure of the Hindi language do not precede any order in the sentences, the dependency relation are observed and introduced between the chunks.

1.7.2 Part of Speech

Generally the group of words called word class have the grammatical properties are assigned to syntax which has similar behavior like the grammatical sentences. Part of speech has another name called lexical class which is concerned with the scientific study of the words. It is used is used to define the morphological and syntactic behavior of the candidates in the text. Hindi language includes noun and verb as linguistic class. In NLP task Part of speech (POS) provide the important information about the words by marking up the words into the document to the particular part of speech [32]. The complete list of POS is defined by the Anncora guidelines [33] to mark the POS tags in Hindi language. The commonly tags are given below:

PRON means pronoun, NNPC means proper noun, PRP means personal pronoun, RB means adverb, NN means noun singular or mass, PREP means preposition, VFM means verb finite main , UNK means out of vocabulary token , VAUX means auxiliary verb, CC means coordinating conjunction , VNN means verb nonfinite nominal.

1.7.3 Syntax

In the scientific language, syntax is defined as the group of the principles and rules that are helpful in making the sentences in the discourse of any language [34]. Mainly it describes the order of the words. In NLP the syntax or rules are used to represent the structure and order of the sentences. Syntax plays a very important role in the process of analysis of language. There are three types of syntax that are Generative, Categorical, Dependency syntax but in this thesis we discussed only two types that are:

1.7.3.1 Categorical grammar or Phrase structure grammar

It is a class of protocols that are used in NLP syntax. It is used the principle of compositionality. It describes the syntax for organizing the sentences which follow the argument-function relation. It uses the combination of basic symbols and inference rules instead of grammar rules for the syntactic structure and this combination of rules are called phrase structure grammar. The combination of rules are iterative in nature that is words are iteratively combined to develop phrases, a large no of phrases are combined to produce sentences. Take an example

e.g. ***Ana ate the noodles with a fork in the evening***

The phrase structure of the above example by using the phrase grammar is give below. In figure1.3 the phrase tree structure sentence S is the root which is terminal symbol and

Childs are non- terminal symbols which are further dived into no terminal symbols.

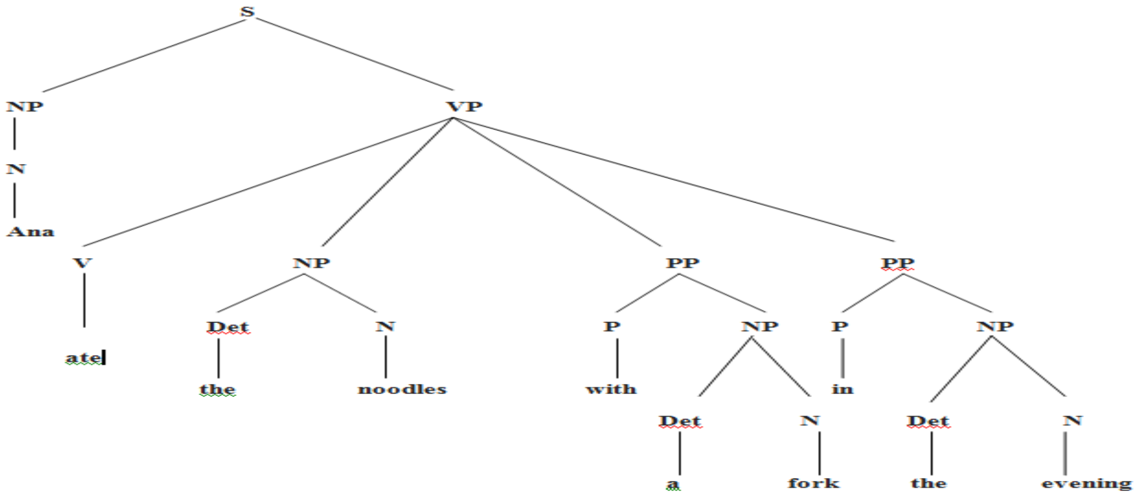


Figure 1.3 Phrase grammar structure

1.7.3.2 Dependency grammar

It is a process of making the structure of the sentences by arranging the syntactic unit preceding the dependency relationship called head- modified relationship. It is contrast to the categorical/phrase structure grammar. In these words of the sentences has the direct dependency links. In process of representing the tree a verb to be the root of the sentence and all the reaming words are indirectly or directly dependent on the root. Take an example e.g. ***Rahul aye the food in evening***

The Rahul is subject, the food is object, ate is verb , with and in are preposition explained in figure 1.4.

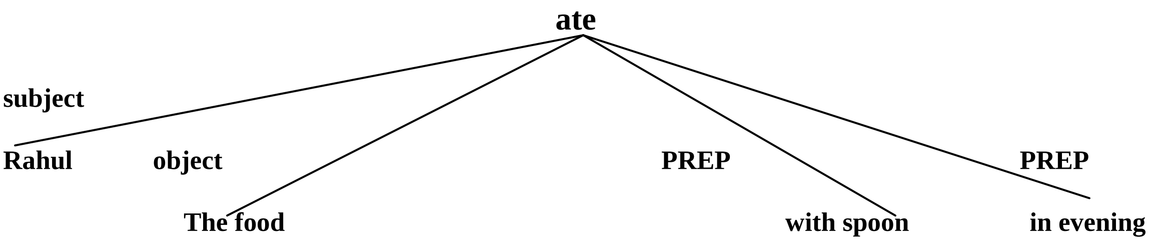


Figure 1.4 Dependency grammar structure

1.7.4 Word Net Library

It is a lexical database that is combination of thesaurus and dictionary. The Hindi WorldNet library is the system for gathering the various different lexical and semantic relations between the Hindi words. Word Net library is also called ontology because it arranges the lexical information of word meaning according to the psycholinguistics principles.

1.8 Hindi Dependency Tree bank Format – AnnCorra , Paninian Grammar and Shakti Standard Format(SSF)

Our goal is to resolve the anaphora resolution over the Hindi Dependency Tree Bank – AnnCorra [35]. We are using this data for our further experiments to resolve the issues of anaphora resolution . We are explaining the tree bank and SSF is brief below:

The full name of AnnCorra is “"Annotated Corpora". The idea of developing the tree bank for Indian languages was first decided at the "Workshop on Lexical Resources for Natural Language Processing", 5 - 8 Jan 2001, held at IIIT Hyderabad. In this format the world knowledge is prepared from the word-formation rules, dictionaries and grammars and schemes for tagging are developed for chunking, sentential analysis, syntactic parsing and POS tagging. In this tree bank the dependency is explained as the Computational Paninian Grammar(CPG) based framework [36] and [37]. In this there are 45 relations or tags that are based on the notation ‘karaka’. The detailed description is given in the Hindi Dependency tag set (http://ltrc.iiit.ac.in/MachineTrans/research/tb/dep-tagset.pdf). This framework has the semantic-syntactic relationship. For tree bank representation Shakti Standard Format (SSF) [38] is used which It has four columns format the first one is for address, the second is for the token, the third is for the category of the node and the fourth is for other features. We have explained the SSF of the sentences in table1.3 and some relationships of tree bank with their meaning are explained below and in table1.2 with the help of example:

e.g. रवि दिल्ली में राहुल के भाई को अपनी कार दे ।

Ravi.ERG delhi.LOC Rahul.GEN brother.ACC his car gave

Ravi told Ana that in delhi he gave his car to rahul’s brother

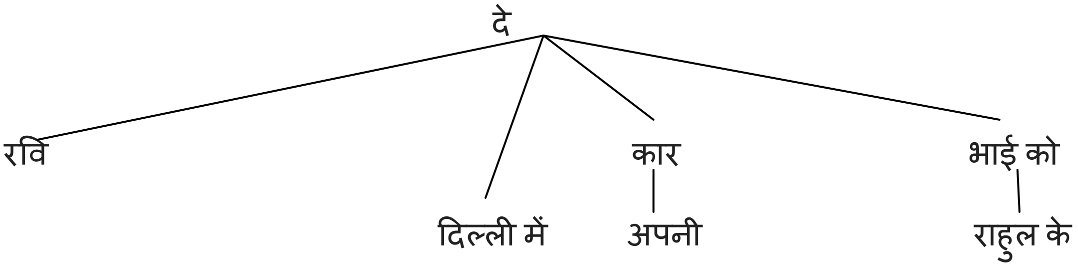


Fig 1.5 Dependency Structure in Hindi

Table1.2: CPG relations

Label	CPG relation	Meaning
k1	Karta	Most independent entity which carries out the action
k2	Karma	The entity on which action is carried out

k4/k4a	Sampradan	Experiencer/receiver
k7p	Apaadan	Location
r6	Sambandh	Genitive/Possessive
Rh	Hetu	Purpose
Rs	Samaanadhikaran	Equivalance
Ccof	Conjunction	Conjunction

Table1.2 shows the explanation of the dependencies relationships and their meaning. In this the ‘karaka’ obeys the semantic – syntactic relationship. Parsing in Hindi can be done by animistic knowledge [39]. Animistic knowledge for Noun phrases is explained in the tree bank . It has three classes First is for human entities called ‘human’, second is for non-human but animate entities called ‘animate’ and third is for inanimate entities called ‘rest’. It also has the feature of NE-Recognizer for Name entity categories in Hindi.

Table 1.3 Shakti Standard format Representation

SrNO	**Token**	**Category**	**Address**	**Features**
1	हिन्दी	NN	0	ROOT
2	संवैधानिक	JJ	3	nmod__adj
3	रूप	NN	0	ROOT
4	से	PSP:से	3	lwg__psp
5	भारत	NNP	8	r6
6	की	PSP:का	5	lwg__psp
7	प्रथम	QO	8	nmod__adj
8		NN	9	Ccof

	राजभाषा			
9	और	CC:और	0	ROOT
10	भारत	NNP	14	r6
11	की	PSP:का	10	lwg__psp
12	सबसे	INTF	14	jjmod__intf
13	अधिक	QF	14	nmod__adj
14	बोली	VM	9	Ccof
15	और	CC:और	9	Ccof
16	समझी	VM	19	nmod
17	जाने	VAUX	16	lwg__vaux
18	वाली	PSP:वाली	16	lwg__psp
19	भाषा	NN	20	k1s
20	है	VM	15	Ccof
21 .	.	.	20	rsym

In Table 1.3 the SSF is represented the sentences. It has four categories that are:

Address: are the human readable tree addresses that are represented by fourth column.

Token: are the actual word in the sentence and groups of words that are represented in second column.

Category or part of speech: are represented by third column (NP, NN, PSP etc)

Others – other user defined features are represented by the last column.

1.9 Organization of the thesis

Hindi language is an independent from the order of the words. It is called ergative language that depends on the verb for resolving the pronouns. It has no proper semantic and syntactic structure like English therefore pronouns in Hindi shows a great deal of uncertainty. To well understand the complexity of the anaphora resolution one has to be clear about the related work in Chapter 2 (Related Work). The research gaps are drilled out according to literature survey of anaphora resolution in Hindi. We formulated our problem and solution in Chapter 3 (Proposed system). In Chapter 4(Experiments and Results) we have computed our results by using the hybrid approach for resolving the gender, number, co-reference and animistic knowledge. Then finally we summarized our thesis in Chapter 5(Conclusion and Future work).

CHAPTER 2- RELATED WORK

In this chapter we have explained the related work on the bases of different algorithms like rule based, learning based, knowledge based etc which we have discussed in chapter 1; section 1.4 in English language.

The problem of the anaphora resolution was first discussed by the Jesperson in1954. [Hobbs, 1978] [12] noticed the anaphora resolution problem in the two nouns refer to the pronoun.

In earlier the anaphora resolution algorithms are based on the heuristics. These algorithms did not use the syntactic constraints. The anaphora resolution algorithm is shifted from the traditional syntax and semantic to the simple semantic and syntax. Some researchers like Bernnar, Friedman and Polard [I9871], Sinder[1981,1983], Grosz, Joshi and Weinstin[l983,1986], and Webber[1988] represent the different versions of the discourse based algorithms. The features like coherence and focusing are used to identify the pronouns.

After 1986 , A system is characterized into two approaches that are Knowledge based and other alternative approaches like Corbonell and Brown [1988] [14] Rich and Luperfiy [1988] and Asher and Wada [1988],. Corbonell et.al presents a varity of syntactic , discourse and semantic factors which combine into the multi-dimentional matric for odering the pronouns.

The other alternative algorithms of anaphora resolution that uses statistical methods where the Basyen condition and probability are used in Sobha L, Patnaik, B.N [1999], Mitkov [1997] [17], Lappin and Lease [1994][19] . In Lappin and Lease [1994] algorithm work on the syntactic representation that is generated by the grammar parser. In this the weights are assigned using the linguistics based ideas. Sobha L and R.N.Patnaik[99] presented the algorithm which assigned the weight to the nouns by using the threshold value and linguists rules . The threshold value is given to the correct items. These items may be non pronominal or pronominal. A comparison is done between the statistical approach and non statistical approach. It is observed that the statistical approach gives the better result in the term of resolving time and accuracy . Kennedy, C and Bogurave, B (1997) [20][21] also used the salience feature for pointing the pronouns. In this only part of speech is done. It did not use the complete syntactic parser.

Rocha [I997], Stuckrdt et.al [1997] used the semantic knowledge for solving the anaphora. Poesio, M et.al. [I997] used the definite description for pointing the pronouns in the text. In this the relationship between the binding description and the definite description is defined and resolved by using bridging description. There is one another anaphora which is called the zero anaphora. Morit, and et.a1.[1997] and Nakaiwa, H [I997] used the pragmatic constraints to resolve the zero anaphora.

There are many other algorithms which used the machine learning approach. Mitkov et.al [I997] [19][26] drafted algorithm which is based on the semi-automatic annotation of pronoun phrases that are pairs in discourse . This algorithm is based on knowledge-poor and robust anaphora resolution method that is followed by post-editing.

There are many systems like CogNIAC [25] which gave high rate of pronouns resolution by using the information of the part of speech tagging, noun phrase recognition, parse tree , basis salience features like number , gender and recency etc, sentence detection etc but it did not resolve the pronouns in the case of ambiguity.

In anaphora resolution a lot of the work is done in other European Languages and English but less amount of work is done in Hindi language. In English Language most of the work is divided into two fundamental approaches that are Rule based and Learning based approaches. In this chapter the related work is divided into three parts. In the first we have provided the idea about the rule and learning based approaches. In the second part we have discussed about the research on basis of dependency structure of Anaphora Resolution in both languages and in third we have discussed the related work on anaphora resolution in Hindi. On the basis of related work the research gaps are also discussed in this chapter.

2.1 Related work on Rule Based and Learning Based methods:

In this part we discuss about the earliest algorithms whichuses various feature to resolve

the anaphora resolution. The Rule based algorithms are of two types Constraint based and preference based. The Learning based algorithms are also of two types Supervised and Unsupervised methods. The brief overview of these approaches is given below:

1. Hobb’s algorithm

It is one of the oldest algorithms which is based on the syntactic constraints to identify the referents of pronoun [40]. The goal is to search the most likely candidates of pronoun in the parse tree of phrase structure of the sentences which contain the pronouns within the sentences or in different sentences. The possible candidates are selected on the bases of c-command and binding theory. The algorithm work as follows:

iv. The search starts from the pronoun node in the parse tree of the current sentence.

v. After that move up in the parse tree until the NP and S node is found. Call this node as ’B’ and the path from the ‘B’ to pronoun as ‘P’.

vi. Now iterate and transverse the path from all the child nodes of ‘B’ that is left to the ‘P’. If the NP node is found which has another NP node that is lying between the path of S and B then NP is act as the referent of pronoun.

vii. If the tree has been reached to B=S then do the left to right breath first transverse of the previous sentence starting with the nearest one.

viii. Now take NP as referent which mostly match with the pronoun.

Now take an example which contains two sentences. The task is to find the pronoun ‘him’ that refer to nouns.

e.g. ***Bob lost the template. John scolded him.***

In Figure 2.1 shows the resolution of pronoun ‘him’ in above example using Hobb‘s algorithm.

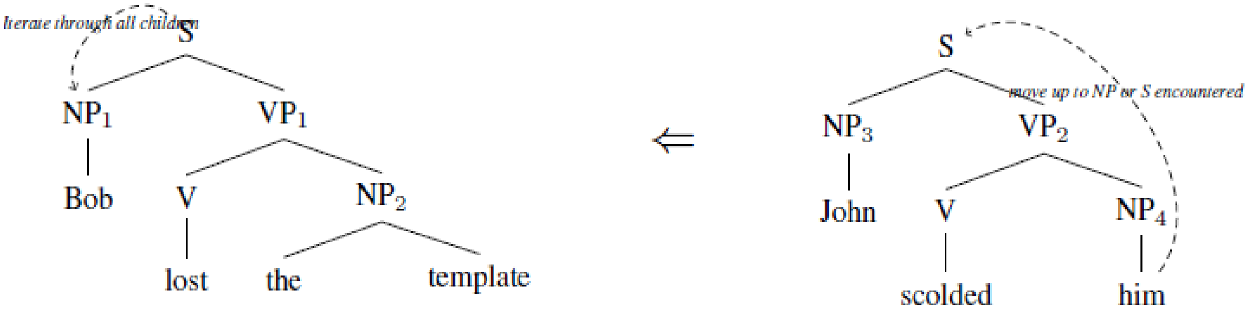


Fig 2.1 Solving the Pronoun using Hobb's algorithm

The algorithm working in the following way :

- The algorithm starts from the pronoun node (NP4) in second sentence.
- After that move up in the tree up-to an 'NP' or 'S' is found. In this case (B=S), only child of 'B' left to path 'P' is 'NP3', but it is rejected since there is no NP node between 'NP3' and 'B' ('S')
- Now the 'S' node is reached, move to the parse tree of previous sentence and traverse in breadth first fashion starting at leftest node 'NP1'
- NP1 is the first node, considered'NP1' as the referent.

2 **Lappin and Leass's algorithm**

To resolve the pronouns this algorithm uses the relative salience of candidate referents. Mainly it focused on the recency, syntax structure, gender and number agreements [41]. In this the candidates are selected on basis of corpus for different languages and domains. The algorithm is work as follows:

- Gather all the possible pronouns.
- Separate the pronouns which do not fulfill the condition of the number and gender agreement.
- Separate the pronouns which do not fulfill the condition of the co-reference constraints.
- Compute the total possible pronouns for the each noun
- Select the noun for pronoun which has higher referent value.

3 **Centering based resolution**

It is based on centers of discourse and the transition between the foci for a coherent discourse [42] . It has the following components:

- Utterance: In this the discourse is divided into single sentences. The set of sentences and clauses are called utterances.
- Center: The entity which is discussed, focused or referred in the text is taken as a center.
- Forward Looking centers: The notation Cf(Un) describes the set of centers that are referred in utterance Un. All the centers in Cf(Un) are ordered on the basis of conditions which act as a filter .These conditions are based on the syntacti rules that is (Subject >Object >Indirect Object >Other)
- Backward Looking center: Every utterance Un has the single backward looking center that is Cb(Un) which is the ordinary elements in the sets of forward looking centers of Un and Un□1. That is Cb(Un) 2 Cf (Un) \ Cf (Un□1). Cb(Un). It is the highest ranked center in Cf(Un□1).
- Preferred center: The Cp(Un) is the greatest ranked center in the Un.

4 **Machine learning approach**

It is the supervised learning approach that is used for resolving the pronouns [43]. They divide the anaphora into two problem classes. The anaphora referents are classified into the pair one is anaphora and second is candidate referents. The features for these pair act as the filter which describes the relation between the noun and the pronoun. The training is given to the discourse for finding out the referent of the anaphora. The classifier is trained according to the rules which are used. This classifier is used to find out the possible candidate according to the rules.

2.2 **Related work using the dependency structure**

In the previous work some approaches use the dependency relations for the anaphora resolution. First one is (Uppalapu and Sharma, 2009) [44] discuss the dependency structure for the Hindi language. This algorithm is the improved version of the S-list algorithm of (Prasad and Strube, 2008) [45]. It used the traditional grammatical relations that is (subject >object >Indirect object >others) for ranking the items in the list and used the Paninian grammatical relations (CPG) i.e. (k1 >k2 >k3 >k4>others). However this algorithm only used the dependency relations as attribute to order the item in the centering based approach. The dependency structure and relation has the syntactic attributes and properties. The dependency structure must be explored to describe the different nouns for different pronouns that are used for resolving the anaphora .

Second one is (Bjrkelund and Kuhn, 2012) [46] describe the dependency structure for English language. They are using the dependency relations as an attribute for the co-reference resolution in a fully learning based approach. The combination of dependency structure and phrase – structure gives better result than using the phrase structure alone.

2.3 Related work on Anaphora resolution in Hindi

Sinha *et al*. presented a translation system of English to Hindi Machine-Aided named as AnglaHindi. Anglabharti was a pseudo-interlingual rule-based translation methodology. It also used example-base and statistics to get more accurate translation for frequently encountered noun and verb phrasals. It used semantics to resolve most of the intrasentence anaphora/ pronoun references. It had problem in making a choice of correct reflexive pronouns .The system generated approximately 90% acceptable translation in case of simple, compound and complex sentences up to a length of 20 words. [47]

Dutta *et al*. presented methods to handle anaphora and used ellipsis and implemented a model by using a prototype of natural language interface (NLI) to databases for Hindi – Matra2. It resolved the issues of Reflexive Pronoun, Possessive Pronoun, and Demonstrative Pronoun where AnglaHindi had problem in making a choice of correct reflexive pronouns. Matra2 was also compare with the Google translator which had also problem in making a choice of correct reflexive , possessive, demonstrative pronoun. It did not differentiate pronouns on gender. [48]

Lakhmani *et al*. presented in paper the report on anaphora resolution for Hindi language. The primary focused on the solution of pronominal anaphora. It also covered the issues related to syntactic and semantic structure of Hindi. It performed a experiment on different kinds of data sets and gave result of approx 71% but only on number agreement and animistic knowledge but Gender agreement don’t show any accuracy.[49]

Lakhmani *et al*. presented the pronominal anaphora resolution for Hindi Language and a computational model for anaphora resolution in Hindi that was based on Gazetteer method which involved many salient factors for resolving anaphora. But the proposed model resolved anaphora by using two factors that is Animistic and Recency. The experiment conducted on different data set - data set 1 - children story – gave accuracy of 65%, data set 2 - news article - gave accuracy of 63%, data set 3 - biography from wikipidea- gave accuracy of 83%.[50]

Chopra *et al.* presented how Anaphora Resolution was useful in performing computation linguistic task in different Natural languages as well in the Indian languages and told about the how the Anaphora Resolution was conducive in handling unknown words in Named Entity Recognition. Transliteration approach was used to solve name entity recognition in various languages.

e.g.: *नुसरत/PER सेब खाती है*

दीपिका/PER कानिपुर/CITY में रहती है

Above, Named Entities in Hindi were transliterated into English as: nusrat, deepika, Kanpur, deepa and nagpur. And it had given approx 96% of result but it did not focus on other issues.[30]

Dutta *et al.* presented the application of Hobbs algorithm for pronominal resolution in Hindi and solved reflexive and possessive pronouns. Modified the Hobb"s algorithm into hobb's naïve algorithm for hindi and do not solve the gender agreement, number agreement , NER etc. [51]

Uppalapu *et al.* presented an algorithm which is in line with S-List (Prasad and Strube, 2000) to resolve the Hindi third person pronouns and showed that there was a refinement of the S-List algorithm in the performance by taking two lists one was present and second was past instead of one. It also explored, how complex sentences could be broken into utterances as motivated by Kameyama (Kameyama, 1997). The algorithm also introduced a another alorithm for resolving the first and the second person pronouns and it had given 61.11%, 77.45% of result on different data sets but did not focus on other issues. [52]

Devi *et al.* presented a system for Indian languages called a generic anaphora engine, which were poor recourses languages. It had analyzed the likeness and unlikeness between the pronouns and their agreement with antecedents. The machine learning approach used the features which could handle major Indian languages. It took shallow parsed text as input and marked Generic Engine and used CRFs, a linear graphical machine learning algorithm to train the system. It had solved the gender problem in past.[53]

Sharma *et al*. drafted approach to resolve Entity-pronoun references in Hindi called hybrid approach and used dependency structures as a source of syntactic information. In this approach, the dependency structure were used a rule-based caliber for resolving the simple anaphoric references and a decision tree classifier was used to solved the more puzzling sentences, using grammatical and semantic features. The results show that by using the dependency structures that gives syntactic knowledge which helps to resolve some specific types of references. Semantic information such as animacy and Named Entity categories further helped to improve the resolution accuracy .It also resolved the Reflexive, Locative, Relative and Personal pronouns. And it had given 70% of accuracy but did not solve the gender agreement and other problem. [54]

Lakhmani *et al*. presented the Gazetteer method for pronominal anaphora resolution for Hindi Language. It had developed a model that used Recency factor which was acted as the baselin and Animistic knowledge which forms the criteria of classification of different pronouns and nouns for performing the pronominal anaphora resolution task for Hindi Language and gave approx 60 to 70% of result but did not solve other issues. [3]

Duttaa *et al.* presented the classification of indirect anaphora in Hindi corpus by using machine learning approach . This was depending on the knowledge of semantic structure provided by the collocation of various patterns and following pronouns were also drilled out. It had given 12.44% result on indirect anaphora in whole data but did not other issues. [55]. The summarization of the anaphora resolution is given in table 2.1.

Table 2.1 Summery of Related Work

Sr. no	**Issues**	**Paper name**	**Work done**	**Limitation**

1	Pronoun resolution (first , second and third person),Number agreement, Animistic	Anaphora Resolution in Hindi Language[49]	1)Resolvedthe issue of pronominal anaphora on the bases of number agreement and animistic knowledge ,added gender agreement 2)It produced result of approx 71%.	1)Gender agreement didn’t show any accuracy. 2)Did not solve NER problem. 3)Other pronouns like reflexive, relative etc were not solved.
2	Pronoun resolution(first , second and third person) , recency and animistic	Anaphora resolution in Hindi Using Gazetter Method[50]	1)Solved the pronominal anaphora resolution using gazetteer method . 2)The proposed model resolved anaphora by using two factors that is Animistic and Recency . 3)It produced result of aprrox 83%	1)Gender agreement didn’t show any accuracy. 2)Did not resolve other pronouns like reflexive, relative etc 3) Number agreement , intrasentential and intersentential, NER problem wer not resolved.
3	Pronoun resolution(first , second and third person, reflexive pronoun), Gender agreement	Resolving Pronominal Anaphora in Hindi Using Hobbs’ Algorithm[51]	1)Solved the pronominal anaphora using hobb’s algo. 2)Focused on reflexive and possessive pronouns, 3)Solved the gender agreement in past	1)Gender agreement didn’t show any result in present. 2)Did not resolve number agreement , intrasentential and intersentential, NER, recency problm.

4	Pronoun resolution(first , second and third person) intrasentential sentences and Number agreement	Pronoun Resolution For Hindi[52]	1)Modified algorithm which was in line with S-List resolving the Hindi third person pronouns in intra sentential It produced 61.11%,77.45% of result on different data sets.	1)It only focused on first ,second and third pronoun not on other pronouns like reflexive,relative etc 2)Did not solve pronouns in intersentential sentences, recency . 3) Gender agreement was not solved.
5	Pronoun resolution(first , second and third person), recency and animistic	Gazetteer Method for Resolving Pronominal Anaphora in Hindi Language[3]	1) Solved the pronominal anaphora by using gazetteer method and used animistic knowledge and recency factor and 3)It produced result of 60 -70%.	1) Did not focus other factors like number, gender agreement. 2)Intrasentential and intersentential, NER, recency problem was not solved.
6	Ambiguities and Unknown words in Named Entity Recognition	Handling Ambiguties And unknown Words In Named Entity Recognition Using [30]	1)It resolved the issue of name entity recognition. 2)Handle the ambiguities using 3)Transliteration approach. 4)It produced result of 90%.	1)Did not focus on other factors like gender agreement, number agreement. 2)Did not resolve other pronouns like reflexive, relative etc.

7	Reflexive, Possessive, Demonstrative, Place, Relative, Indirect pronounces resolution **.**	Anaphora Resolution in Hindi: Issues and Challenges[48]	1)Matra2 is compared with the, Anglahindi, Google translator which had also problem in making a choice of correct reflexive, possessive, demonstrative pronoun. It solved the issues.	1) It did not differentiate pronouns on gender. 2) NER problem was not resolved. 3) Did not resolve number agreement, recency.
8	Gender Agreement.	A Generic Anaphora Resolution Engine for Indian Languages[53]	1) Developed the generic algorithm for Indian languages by using the machine learning approach for the sameness and variations between the various pronouns and their agreement with their antecedents	1) Did not focus on number agreement, NER, recency. 2) Had used very minimal resources
9	Reflexive, Possessive, Demonstrative, Place, Relative, Indirect pronounces resolution ,Name entity Reorganization, recency and animistic	A Hybrid Approach for Anaphora Resolution in Hindi [54]	1) Resolved Entity-pronoun references in Hindi by using dependency structures with rule-based module 2) Worked on Reflexive, Locative, Relative and Personal pronouns, animacy , name entity recognition	1) Did not focus on other 2) salient factors only focused resolving only entity pronouns
10	Reflexive,	Machine	1) Presented	1) Only focused on

	Possessive, Place, Relative, Indirect ,Demonstrative pronounces resolution	Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items[55]	classification of indirect anaphora in Hindi corpus by using machine learning approach 2)It produced result on 12.44% of indirect pronouns .	indirect anaphora. 2) Did not solve other issues.
11	Handel IntrasententiaI and Intersentential sentences	Angla Hindi: An English to Hindi Machine-Aided Translation System[47]	1) Developed AnglaHindi System. 2) It used semantics knowledge to resolve most of the intrasentence pronoun references. 3) It produced result of 91% on simple, compound,complex sentences up to 20 words.	1) It had problem in making a choice of correct reflexive pronouns . 2) Did not focus on other issues.

2.4 Research Gaps

It can be observed from the table 2.1 that the area of Anaphora Resolution is still lacks of resolving various issues. The research gaps of existing approaches are listed below:

1) Very less amount of work has been done on Gender Agreement.

2) Researchers should develop better algorithm to handle pronouns in intrasentential, intersentential, entity and event sentences.

3) Standardization of single text processing tool for anaphora resolution is required by limiting the utilization of corpus.

4) There is need to develop a better algorithm which can work on all issues at the same time and it is also required researchers to develop approaches for Indian languages which resolve anaphora resolution.

5) Researchers used only one metric that is F value to measure the performance. But there are many other metric for calculating the performance of anaphora resolution are ,

ware BLANC, MUC, B3, CEAF and these metric exhibits significantly different behaviors from each other and globally accepted .

6) Very less amount of work has been done on anaphora resolution by using the dependency structure.

2.5 Summary

It show the brief summary of the chapter .In this we have mentioned previous research on the anaphora resolution. The related work of anaphora resolution in English language is described on the bases of different algorithms like rule based, learning based, knowledge based etc. we have provided the idea about the rule and learning based approaches like Hobb's algorithm , center based and lappin and leass's algotithm and also discussed about the research on basis of dependency structure of Anaphora Resolution in both languages . we have drilled out the not the used of all evaluation metrics etc on the basis of the related work on anaphora resolution in Hind

CHAPTER 3- PROPOSED SYSTEM

We have discussed various research gaps in Chapter 2, section 2.4. By observing these research gaps we have derived a problem of anaphora resolution in Hindi language.

3.1 Problem

A hybrid approach for Anaphora resolution in Hindi to resolve Gender, Number, Co-reference, Animistic Knowledge.

3.2 Problem Description

The task of this thesis is to resolve the anaphora resolution in Hindi which refers to the task of identifying mentions (basically noun phrases) that denote the same real world objects, or entities. Many crucial applications involving Natural Language Processing (NLP), for example, Information extraction, question-answering, machine translation, text summarization etc. which we have already discussed in chapter1 that require the task of pronoun resolution to be performed. Because Hindi is a verbal as well as free order word language. The aim is to utilize anaphora resolution using a hybrid approach of multi-objective feature selection and rule based using gender, singularity, co-reference, animistic knowledge. We also used Hindi dependency structure Anncora to resolve the pronoun resolution. We trained our system with the word net library in this we use learning approach to train the machine.

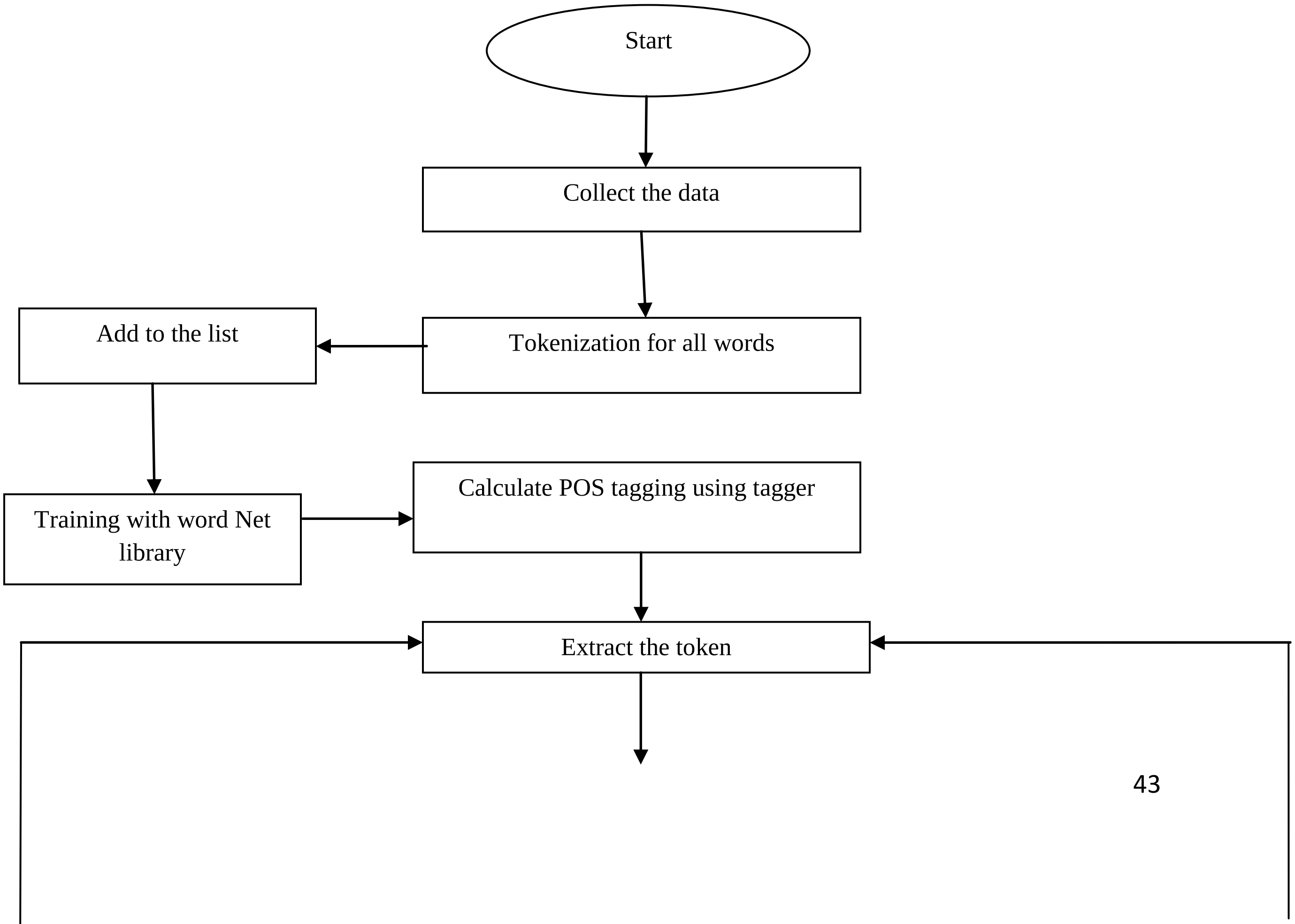
And also there does not exist any globally accepted metric for measuring the performance of anaphora resolution, and each of F-value MUC, B3, CEAF, BLANC exhibits significantly different behaviors. System optimized with respect to one metric often tends to perform poorly with respect to the others, and therefore comparing the performance between the different systems becomes quite difficult. In our work we determine the most relevant set of features that best optimize all the metrics.

3.3 Our Approach

In this we use hybrid multi-objective feature selection, rule based approach, Hindi

dependency structure (CPG) and Learning approach (system training) to resolve the Gender, Number agreement, Recency factor and Animistic knowledge.

3.3.1 Flowchart of the System


Start
Collect the data
Add to the list
Tokenization for all words
Training with word Net library
Calculate POS tagging using tagger
Extract the token

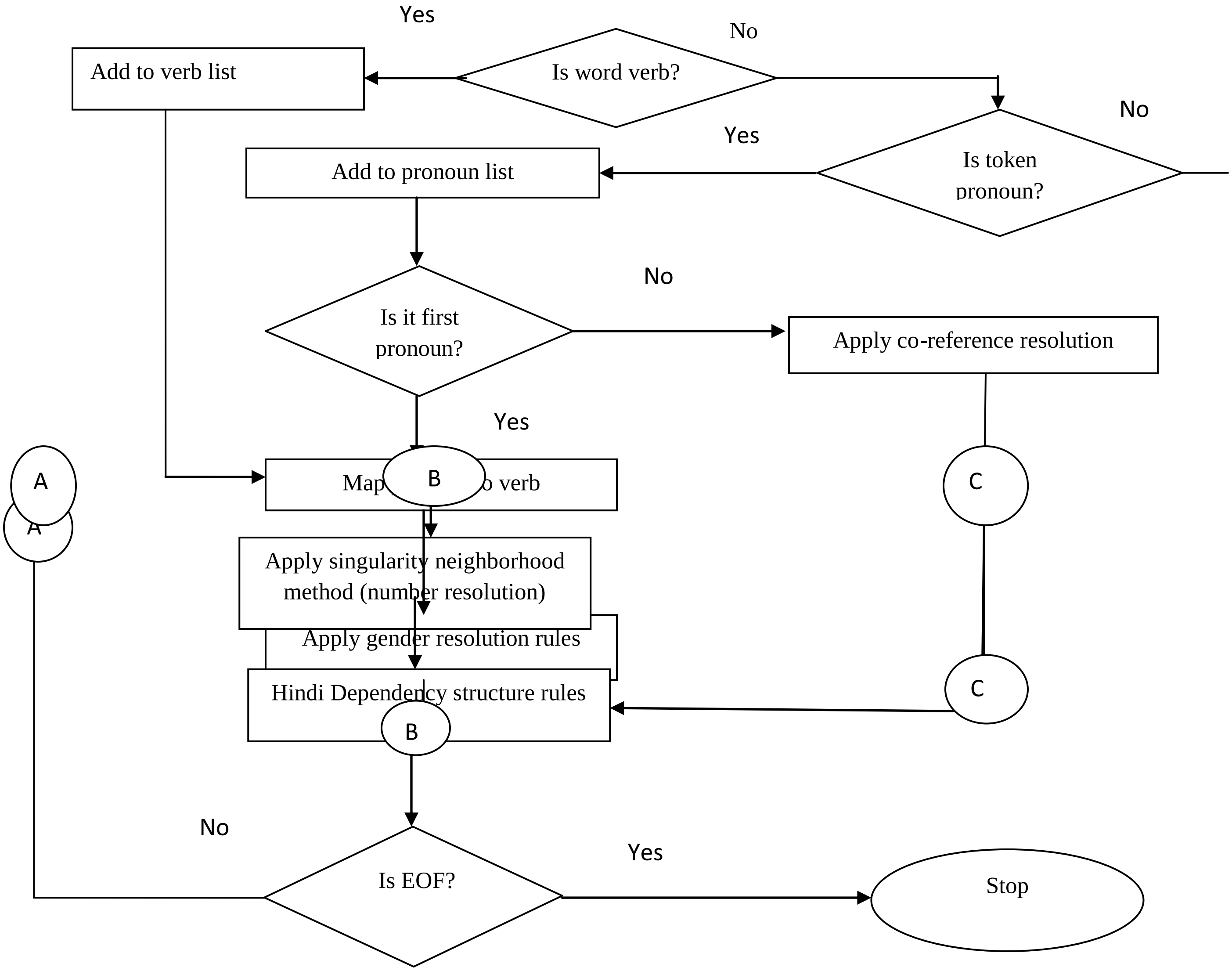

Yes
No
Add to verb list
Is word verb?
No
Yes
Add to pronoun list
Is token pronoun?
No
Is it first pronoun?
Apply co-reference resolution
Yes
A
A
Map
B
o verb
C
Apply singularity neighborhood method (number resolution)
Apply gender resolution rules
Hindi Dependency structure rules
C
B
No
Is EOF?
Yes
Stop

Figure3.1: Flowchart of the System

The working of the proposed system is shown in fig 3.1

1. The data is taken as input than it is chopped into pieces the process is known as tokenization.

2. Then the tokens are added into the list where it will be trained with word net library so that we can calculate part of speech (POS) tagging which marks the word in text corresponding to the particular part of speech.

3. Than the token is extracted from the list to check if it is a verb than it will be added to the list of verb and if not than it is a pronoun than it will be added to pronoun list.

4. If the token is not pronoun than it will again extract the token from the list.

5. Now if the pronoun is a first pronoun than it will be mapped to the verb. Than gender agreement rule is applied to resolve the pronoun based on gender and the number agreement rule is also applied which will resolve the pronoun based on number.

6. If it is not the first pronoun than the co-reference resolution rule is applied.

7. If all the pronouns are resolved than it will end the task and if not than it will again extract the tokens from the list.

3.3.2 System Design

The design of the system is describe in figure 3.2. It has followed the same basic process that is described in chapter in section 1.6. First the text is auto identified by machine. Then it is broken into pieces by removing the special characters, stop words called token. These tokens are added to the word list. The POS tag is calculated by using the Word Net library according to part of speech two lists are generated one is verb list and second is pronouns list. The rules are applied to the both list and after the filtration the output is generated on the bases of gender, number, co-reference and animistic knowledge.

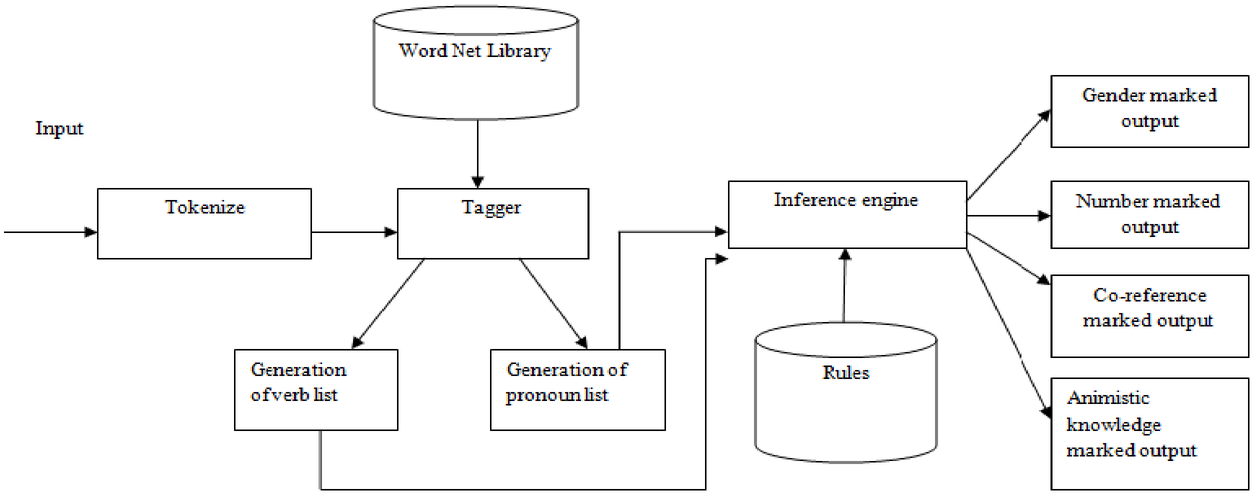


Figure 3.2: System Design

3.3.3 Algorithm

We have implemented the algorithm of the system described in figure 3.3. In our algorithm the data set is taken as the input. The data is preprocessed which include the Removal of special character: like '.' ; '<'; '>' ; ',' ; '_' ; '-' ; '\\' ; '(' ; ')' ; ':' ; ';' ; '[' ; ']' ; '{' ; '}' or '/':, removal of stop words that refer to the most common words in a language, word separation, tokenization that is the task of chopping it up into pieces, called tokens, stemming that refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes, case conversion: change lower case to upper case or vice versa. After that word list is created and POS tag is calculated. According to the part of speech the verb and pronoun list is generated and rules are applied to these lists after the filtration of the lists the gender, number, co-reference and animacy marked output is generated.

Input: Text Document

Output: Gender, number , co-reference, animistic Marked Output

Algorithm:

```
1. Set i to 1 // For each sentence in the document
2. Removal of special character.
3. Removal of stop words
4. Word separation
5. Tokenization
6. Stemming
7. Case conversion
8. Create wordlist
9. Selection of possible candidate
10. POS tagging
11. Apply linguistic rules for gender agreement
      i.   If the gender of the verb contains ‘a’/ ‘◌ा ‘at the last then the noun is
           marked as masculine gender.
      ii.  If the gender of the verb contains ‘e’/ ‘◌ी ‘at the last then the noun is
           marked as feminine gender.
      iii. If the gender of the verb contains ‘a’/ ‘◌ा ‘at the last then the pronoun is
           marked as masculine gender.
      iv.  If the gender of the verb contains ‘e’/ ‘◌ी ‘at the last then the pronoun
           is marked as feminine gender.
12. Apply linguistic rules for  Number agreement
      i.   If the word of the verb contains ‘ae’/ ‘◌े ‘ and ‘aae’/ ‘◌ै’at the last then
```

```
14. Apply CPG rules
        i.    (k1 >k2 >k3 >k4>others)
15. Marked out put
16. i ++
17. If (i< n ) go to step 2 else end.
```

Figure 3.3: Proposed Algorithm of the System

3.3.4 Rules for Gender and Number Agreement

We define rules for gender and number agreement on the basis of verb to resolve the anaphora resolution in figure 3.4. The rules are given below:

Rule 1:- If the gender of the verb contains ‘a’/ ‘◌ा ‘at the last then the noun having a masculine gender.

Rule 2:- If the gender of the verb contains ‘e’/ ‘◌ी ‘and ‘ई’ at the last then the noun having a feminine gender.

Rule 3 :- If the gender of the verb contains ‘a’/ ‘◌ा ‘at the last then the pronoun having a masculine gender.

Rule 4 :- If the gender of the verb contains ‘e’/ ‘◌ी ‘ and ‘ई ’at the last then the pronoun having a feminine gender.

Rule 5 :- If the word of the verb contains ‘ae’/ ‘◌े ‘ and ‘aae’/ ‘◌ें’at the last then the noun/ pronoun is plural .

Rule 6:- If the word of the verb contains ‘au’/ ‘‘◌ु “ , ‘e’/ ‘◌ी ‘ ‘and ‘a’; ‘◌ा’ at the

Figure 3.4: Generated Rules for Gender and Number Agreement

The explanation of the rules is given below:

In Rule 1:- If the gender of the verb contains ‘a’/ ‘ा ‘at the last then the noun having a masculine gender. E.g. ” राहुल बाज़ार जा रहा था |”

Now in this the verb” रहा” ends with “ा” that means Rahul is a masculine gender.

In Rule 2:- If the gender of the verb contains ‘e’/ ‘ी ‘ and ‘ई’ at the last then the noun having a feminine gender. E.g. “मीना घर चली गयी |”

Now in this verb “चली “ ends with “ी ” that means Meena is a feminine gender. Similarly it applies to other rules.

Rule 3 :- If the gender of the verb contains ‘a’/ ‘ा ‘at the last then the pronoun having a masculine gender. E.g. “राहुल बाज़ार जा रहा था और वहाँ वो मीना से मिला | “

Now in this verb “मिला” ends with “ा” that means “वो ” pronoun having a masculine gender. Rule 4:- If the gender of the verb contains ‘e’/ ‘ी ‘ and ‘ई ’at the last then the pronoun having a feminine gender. E.g. “मीना सब्जी लेने आई थी | वो फिर घर चली गयी | “

Now in this verb “चली ” ends with “ई” that means “वो ” pronoun having a feminine gender. Rule 5:- If the word of the verb contains ‘ae’/ ‘◌े ‘ and ‘aae’/ ‘◌ै’at the last then the noun/ pronoun is plural . E.g . “राहुल और मीना चले गये हैं । वे बाज़ार नहीं गये हैं।“

Now in this verb “गये ” ends with “◌े” “that means “वे” pronoun refers to “राहुल और मीना ” noun is plural.

Rule 6:- If the word of the verb contains ‘au’/ ‘‘◌ू “ , ‘e’/ ‘◌ी ‘ ‘and ‘a’; ‘◌ा’ at the last then the noun/ pronoun is singular E.g. “मैं काम कर रहा हूं | “मैं काम कर रही हूं |”

Now in this verb “रहा हूं ” ends with “◌ू” , “◌ा” that means “मैं” pronoun refers to masculine singular noun. In second example the verb “रही हूं” ends with “◌ू” , “◌ा” that means “मैं” pronoun refers to feminine gender.

3.3.5 <u>CPG Rules</u>

The CPG rules are known as computational Paninian grammatical relations or rule [35] like (k1 >k2 >k3 >k4>others) that we described in the chapter 1, section 1.8. The some rules are explained below in table 3.1:

Table 3.1: Explanation of Paninian grammatical relations

Sr No.	Tag Name	Tag description	Example
1.1	k1	karta (doer/agent/subject) *Karta* is defined as the 'most independent' of all the *karakas* (participants).	(1) **rAma** bETA hE (2) **sIwA** KIra sAwI hE
1.2	pk1	prayojaka karta (Causer) The causer in a causative	(1) ***mAz ne*** *bacce ko KanA* ***KilAyA***

		construction.	
1.3	jk1	prayojya karta (causee) The causee in a causative construction.	(1) *mAz ne AyA se* ***bacce ko*** *KAnA KilavAyA*
1.4	mk1	madhyastha karta (mediator-causer) The mediator-causer in a causative construction.	*(1) mAz ne* ***AyA se*** *bacce ko KAnA KilavAyA*
1.5	k1s	vidheya karta (karta samanadhikarana) Noun complements of *karta*	(1) *rAma* ***buxXimAna*** *hE* (2) *xaniyA iwanI* ***vyavahArakuSala*** *na WI*
2.1	k2	karma (object/patient) *Karma* is the locus of the result implied by the verb root.	(1) *rAma rojZa* ***eka seba*** *KAwA hE* (2) *rAma ne bAjZAra meM* ***ravi ko*** *xeKA*
2.2	k2p	Goal, Destination The destination or goal is also taken as a *karma. k2p* is a subtype of *karma* (k2). The goal or destination where the action of motion ends is a k2p.	(1) *rAma* ***Gara*** *gayA* (2) *rAma ko* ***xillI*** *jAnA padZA*
2.3	k2g	gauna karma (secondary karma)	(1) *ve loga gAMXIjI ko* ***bApU*** *BI kahawe hEM*
2.4	k2s	karma samanadhikarana (object complement) The object complement is called as *karma samanadhikarana.*	(1) *rAma mohana ko* ***buxXimAna*** *samaJawA hE*
3	k3	karana (instrument)	*(1) rAma ne* ***cAkU***

		karana karaka denotes the instrument of an action expressed by a verb root. The activity of *karana* helps in achieving the activity of the main action.	***se*** *seba kAtA* (2) *sIwA ne* ***pAnI se*** *GadZe koBarA* (1)
4.1	k4	sampradaana (recipient) Sampradana karaka is the recipient/beneficiary of an action. It is the person/object for whom the karma is intended.	*(1) rAma ne mohana ko Kira xI* *(2) rAma ne hari se yaha kahA*
4.2	k4a	anubhava karta (Experiencer) The experiencer/perceiver in perception verbs such as seems, appear, etc..	*(1) muJako rAma buxXimAna lagawA hE* *(2) muJako cAzxa xiKA* *(3) rAma ko BUka lagI*
5.1	k5	apaadaana (source) apadana karaka indicates the source of the activity, i.e. the point of departure. A noun denoting the point of separation for a verb expressing an activity which involves movement away from is apadana.	*(1) rAma ne cammaca se katorI se Kira KAyI* *(2) cora pulisa se BagawA hE*
5.2	k5prk	prakruti apadana (‘source material' in verbs denoting change of state) A special case of apadaan i.e k5. This is because there is a conceptual	*(1) jUwe camade se banawe hEM*

		separation point from the original raw material ‘camade’ (leather) to the finished product ‘jUte’ (shoes). The two states in this change of state action are referred to as prakriti 'natural' and vikruti 'change'.	
6.1	k7t	kaalaadhikarana (location in time) Adhikaran karaka is the locus of karta or karma. It is what supports, in space or time, the karta or the karma.	*(1) rAma xilli meM rahawA hE* *(2) bacapana meM vaha bahuwa SEwAna WA*
6.2	k7p	deshadhikarana (location in space) The participant denoting the location of karta or karma at the time of action.	*(1) mejZa para kiwAba hE* *(2) havA meM TaMdaka hE*
6.3	k7	vishayaadhikarana (location elsewhere) Location other than time and place.	*(1) ve rAjanIwi para carcA kara rahe We* *(2) harI ne svawanwrawA saMgrAma meM hissA liyA*
7	k*u	saadrishya (similarity) This can be used for marking both similarity and comparison. The tag is marked on the comparand in a comparative construction. Since the compared entity can compare with any karaka, the tag includes a star. '*'	*(1) [[k1u]] rAXA mIrA jEsI sunxara hE* *(2) [k2u] sIwA mIrA ko rAXA jEsI sunxara mAnatI hE*

		in the tag name is a variable for whichever karaka is the comparee of the comparand. Therefore, while marking the comparand (the compared entity), the * would be replaced by the appropriate karaka label.	
8.1	r6	shashthi (possessive) The genitive/possessive relation which holds between two nouns.	*(1) sammAna kA BAva* *(2) puswaka kI kImawa*
8.2	r6-k1, r6-k2	karta or karma of a conjunct verb (complex predicate)	*(1) [r6-k1] kala manxira kA uxGAtana huA* *(2) [r6-k2] manwrIjI ne kala manxira kA uxGAtana kiyA*
8.3	r6v	('kA' relation between a noun and a verb) Instances where a noun with 'kA' is attached to the verb but does not have any *karaka* relation. Instead, it does indicate a sense of possessesion.	*(1) rAma ke eka betI*
9	adv	kriyaavisheshana ('manner adverbs' only) Adverbs of manner are marked as 'adv'. Note that the adverbs such as	*(1) vaha jalxI jalxI liKA rahA WA* *(2) vaha bahuwa*

		place, time, etc. are not marked as 'adv' under this scheme.	*wejZa bolawA hE*
10	sent-adv	Sentential Adverbs Adverbial expressions that have entire sentence in their scope.	*(1) isake alAvA, BakaPA (mAovAxI) ke rAmabacana yAxava ko giraPZawAra kara liyA gayA*
11	rd	prati (direction) The participant indicating ‘direction’ of the activity.	*(1) sIwA gAzva kI ora jA rahI WI* *(2) rAma ke prawi mohana ko SraxXA hE*
12	rh	hetu (cause-effect) The reason or cause of an activity.	*(1) mEne mohana kI vajaha se kiwAba KArIxI*
13	rt	taadarthya (purpose) The purpose of an action.	*(1) mEne mohana ke liye kiwAba KArIxI* *(2) mohana padZane ke liye skUla jAwA hE*
14.1	ras-k*	upapada__ sahakaarakatwa (associative) Two participants performing the same action but syntactically one is expressed as primary and the other as its associate, the associate participant.	*(1) rAma apane pIwAji ke sAWa bAjZAra gayA*
14.2	ras-neg	Negation in Associatives	*(1) rAma pIwAjI ke*

			binA gayA
15	rs	relation samanadhikaran (noun elaboration) Elements (normally clauses) which elaborate on a noun/pronoun.	*(1) bAwa yaha hE ki vo kal nahIM AyegA*
16	rsp	relation for duratives The durative expressions have two points – a point of starting and an end point. The expression as a whole may express time, place or manner etc. The tag 'rsp' shows the relation between the starting point and the end point of a durative	*(1) 1990 se lekara 2000 waka BArawa kI pragawi wejZa rahI*
17	rad	Address words Terms such as SrImAnajI, paMdiwajI etc. are the address terms.	*(1) mAz muJe kala xillI jAnA hE* *(2) mAstara sAhaba, kyA kala skUla KulA hE*
18	nmod__relc, jjmod__relc, rbmod__relc	Relative clauses, jo-vo constructions	*(1) merI bahana [jo xillI meM rahawI hE] kala A rahI hE*
19	nmod	Noun modifier (including participles) An underspecified relation employed to show general noun modification without going into a finer type.	*(1) pedZa para bETI cidZiyA gAnA gA rahI WI*
20	vmod	Verb modifier Another underspecified tag. For some relations getting into finer subtypes is	*(1) vaha KAwe hue gayA*

		not yet possible. Such relations are annotated with slightly underspecified tag.	*(2) vaha KAnA Kakara gayA* *(3) rAma sAzpa ko xeKakara dara gayA.*
21	jjmod	Modifiers of the adjectives	*(1) halkI nIlI kiwAba* *(2)*
22	pof	Part of relation Part of units such as conjunct verbs.	*(1) rAma ravi kI prawIkSA kara rahA WA.* *(2) rAma ne eka praSna kiyA* *(3) sadZaka cOdZI huI*
23	ccof	Conjunct of relation Co-ordination and sub-ordination.	*(1) rAma seba KAwA hE Ora sIwA xUXa pIwI hE* *(2) rAma ne SyAma se kahA ki vaha kala nahIM AyegA*
24	fragof	Fragment of	*(1)BAkaPA (mAovAxI) ke* *(2) giraPZawAra kara liyA gayA*

25	enm	Enumerator	*(1) 1. Apa apanA kara samaya se xe sakawe hEM*

There are total 42 tags that are used to resolve the anaphora resolution. In our work we also use these tags.

3.3.6 Pseudo Code for Co-reference

The algorithm for calculating the co-reference is explained in figure 3.5. In this higher weight is assigned to noun which is closest to pronoun. . In this process first the training is given to the system by library Hindi keywords and dependency tagger. The POS is calculated by using POS tagger. After calculating the part of speech of the words verb and pronoun lists are generated than check the list if the word is pronoun than the higher

weight is assigned to the nearest noun.

I: Input file fi containing Hindi text

Li: Library of Hindi Keywords

POS: Part of Speech Tagger for each word

Di: Dependency tagger for each word in each sentence containing tag name Tn and tag description Td
Wj : Weight of the word
ccof : Conjunct of relation

```
for each line in I
    Store line Lj
    Obtain wj by tokenizing line Lj
end for
Train the algorithm using library Li
Train Dependency tagger Di with file I
for each word wj in I
    Find POSw using POS tagger
    Use Di for obtaining Tnw and Tiw
    if Tnw equals k1
        Assign weight wk to word wj
        Obtain POS of word wj
        if word wj is pronoun
            Check for weight of nearby noun with Tnw equals k1
            Find verb in vicinity and assign weight
```

Figure 3.5: Pseudo Code for Co-reference

3.4 **Summary**

In this the brief summary of the chapter is described. Firstly we described our problem of the thesis on the basis of the research gaps. We implemented our system using the rule based and learning based approach called hybrid approach for resolving the pronouns based on gender, number, co-reference and animistic knowledge. The verbs are considered as important entity in generated the heuristics rule for gender and number agreement. The higher weights are assigned to nouns which are closest to pronouns to resolve co-reference problem. The training is given to the system by word Net library, POS tagger to calculate the part of speech of the word with the help of this lists are generated and dependency tagger and CPG rules are also used to train the system for resolving the anaphora.

CHAPTER 4 – EXPERIMENT AND RESULTS

We have implemented our system using Python 3.4.3 with NLTK toolkit. In the implementation of system we have used the algorithms which we have described in chapter 3, section 3.3. In this section we are discussing our experiment and result on the three different dataset. We have implemented two different approaches one is rule based and other is Learning based approach. In rule based approach we use heuristics rule and CPG rules for finding the gender, number, co-reference, animistic based pronouns. These rules act as the filter for the candidates and in Learning based approach we use learning techniques to train our system so that it can calculate POS that is based on the available knowledge for Hindi in the Word Net library. The combination of this approach is called Hybrid approach. In this we use very limited rules of the Hindi dependency tree bank for explaining the dependency and animistic knowledge. This chapter is categorized into the datasets to be used in our research work in section 4.1, section 4.2 explained the features , the evolution metrics are described into section 4.3 and results are illustrated into the section 4.4.

4.1 Datasets to be used in our research work

Based on the earlier works done in this field, it is difficult to get a comprehensive view of the research on anaphora resolution related to Indian languages because each of these was developed using the self-generated datasets. We also used the self- generated dataset. Our data set is divided into three parts first dataset is generated from the children story domain (http://abhivyakti-hindi), second dataset is generated from the news article domain (http://webduniya//hindi_news) and third one is from bibliography articles domain. We implemented our approach by using three different datasets. This has different level of pronoun complexity. To check our efficiency of the system we combined the different articles of single domain into one dataset so that the complexity of the pronoun resolution is increase.

4.2 Features

For this Hybrid approach we used some features that are describe below:

- Pronoun: In linguistic language pronouns are the proxy of the nouns. Pronouns itself is the feature because there is not any information is available. The pronouns play a very
  important role in resolving the anaphora resolution.
- POS tag: It is the part of speech tag that assign part of speech of text in any language to each word that are noun, verb , adverb, adjective etc. For tagging purpose a piece of software is used called part of speech tagger [56][57]. It is consider as the feature because the head of the candidate is more often occurs as the ‘NN’.
- Distance: In this we calculate the distance between pronoun and more often occurring noun. The distance between the anaphora and entity of ‘NN’ is considered as the feature. It includes the referent entity, number of sentences between the pronouns, reorganization of the chunks that contains NP phrases.

- Term frequency and Weighting: It is used as the feature. It is used to calculate the raw frequency of the term means how many time the term is occur in the document. The weight is given to that entity which is closer to that term.
- Mention detection: It is the process of recognize the expression that may be pronouns or nouns in the anaphora resolution. The tokens are selected as pronoun which has the tag ‘PRP’ after the POS tagging and the pronouns are added to the list. The tokens are selected as verb which has the tag ‘VFM’ and add to the verb list. The referents are detected according to the rules. The following rules gives the verification of anaphora and its antecedent:
    - Consider the entire NP token as the mention entity.
    - Select the NP token which has head as Pronoun as anaphora.
    - Reject the pronoun which is indefinite.
- Resolving Algorithm: After the selection of the pronouns, the entities are checked whether a particular entity referent is referent to a given pronoun with the help of classifier. The classifier is trained according to the applied rules.

4.3 Evaluation metrics

There is no generic performance metric exists, there are a lot of metrics used by several researchers. There does not exist any globally accepted metric for measuring the accuracy of anaphora resolution, and each of MUC, B3, CEAF, BLANC and F-value exhibits significantly different behaviors. System optimized with respect to one metric often tends to perform poorly with respect to the others, and therefore comparing the performance between the different systems becomes quite difficult. In our work we determine the most relevant set of features that best optimize all the metrics. Here we develop an efficient technique by selecting the various features in anaphora resolution to determine the best evaluation metric for evaluating the anaphora resolution. We also use the Fleiss’s Kappa metrics to calculate the agreement. The detailed description of all evaluation matrices is given below [58] [59]:

4.3.1 Notations

- There are two systems one is called GOLD system in which the pronouns are detected and resolved by the human and second one is called SYS system in which the pronouns are detected and resolved by the machine.
- Singleton mention refers to the noun which occurs only single time after resolving the pronouns by the system.
- Doubleton mention refers to the nouns which occur two times after pronoun resolution in the system.
- Multiple mentions refer to the frequency of the nouns after the pronoun resolution in the system.
- After resolving the pronouns by the machine is called response chain and these predicted pronouns pointing to the noun called key chain
- False positive (FP) is refer to the pronoun mentions to entity is false in response chain
- False negative (FP) means the entity according to key chain but are something else in response chain.
- S (d) is defined as the records detected by the system.
- S1 (d) is the correct or relevant records detected by the system.
- K (d) is defined as the relevant records in the document.
- N is the total no of mentions, specifically,

- $S(d) = \{S_j : j = 1, 2, \cdots, |S(d)|\},$ Equation (4.1)
- $S1(d) = \{S1_x : x = 1, 2, \cdots, |S1(d)|\},$ Equation (4.2)
- $K(d) = \{K_k : i = 1, 2, \cdots, |K(d)|\},$ Equation (4.3)

Where $K_k$is series of mentions in K(d), $S1_x$ is series of mentions in S1(d) and S_j is series of mentions in S(d) respectively.

i. **MUC**

It is the link based evolution matrix called message understanding co-reference [59] that calculates the minimum link between the mention (GOLD and SYS). To calculate the recall (R), the total no of links between GOLD and SYS is divided by the minimum number of links that are required to specify GOLD. To calculate precision (P), this number is divided by the minimum number of links that are required to specify SYS [60]. The process of calculation the MUC is given below:

- Find the number of singleton mentions in the document (d).
- Find the number of doubleton mentions in the document.
- Find the number of multiple mentions in the document.
- Find the precision (P) = $\sum_{1}^{N}(S1\backslash S) \setminus N$ Equation(4.4)
- Find the recall (R) = $\sum_{1}^{N}(S1\backslash K)\backslash N$ Equation(4.5)
- Find the MUC F-score = $\frac{2\ P * R}{(P+R)}$ Equation(4.6)

ii. **B3**

In this, to obtain recall all the intersecting mentions between the SYS and GOLD is calculated and dived by the total number of mention in the GOLD [61]. To obtain the precision the number of joining mentions between the SYS and GOLD is calculated and dived by the total no of mentions in the SYS. The process of calculating the B3 is given below:

- Calculate the B3 precision

  Singletons are ignored for false positive while consider for false negative. Calculate the B3(P) by using equation 4.4.
- Calculate the B3 recall

  Singletons are considered for FN. B3 (R) is calculated by using equation 4.5.
- B3 F-score is calculated by using the equation 4.6.

iii. **CEAF**

It is the best method to compute the one to one mapping between the entity in the SYS and GOLD which means each SYS entity is mapping to at most single GOLD entity [62]. It maximizes the similarity of the best mapping. The score of recall and precision are same when true mentions are involved. In both the common mentions between every two mapped entities is divided by the no mentions. The process of calculating the CEAF is given below:

- In response discard all singletons and doubletons

  In original

  if Precision

  Keep doubletons

  If recall

  Discard doubletons

- Find the precision (P) using equation 4.4.
- Find the recall (R) using equation 4.5.
- Find the CEAF F-score using equation 4.6.

iv. **F – score**

The F- score combines precision and recall [63]. RECALL is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage. PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage. The process of finding the F-score is given below:

- Find the recall (R) = $S1\backslash S$ Equation(4.7)
- Find the precision (P) = $S1\backslash K$ Equation(4.8)
- Find the F-score using the equation 4.6.

5 Fleiss's Kappa

It is the statistical evaluation matrix which is used to check the reliability of the agreement over the multiple raters [64]. P(A) is the proportion of the time the judges agreed. P(E) is the proportion of the time they would be expected to agreed by chance. The strength of the agreement is described in table 4.1.

- For number of judges that assign category j to pronouns i = xij
- The Fleiss's kappa is calculated as :

  $$k = \frac{P(A)-P(E)}{1-P(E)} \qquad \text{Equation (4.9)}$$

- The numerator of equation indicates the degree of agreement that is attainable above chance.

- The denominator indicates the degree of agreement actually achieved above chance.

Kappa statistic	Strength of agreement
<0.00	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.0	Almost perfect

Table 4.1 Strength of agreement

4.4 Results

We have calculated our results by using various globally existing metrics that are explained in chapter 4, section 4.3. We have tested our approach on three different types of data sets that are new story and bibliography articles. We calculated our result in two parts, first part we have used the rule based approach for finding the gender and number agreement and in second part we have used rule based , learning based, CPG dependency rules two solve the four issues of anaphora resolution that are gender, number agreement, co-reference and animacy. The results are shown below:

4.4.1 Results Based on Gender and Number Agreement

We have tested our rule based approach on gender and number agreement. We have taken different data set, one is news article and other is story article. Based on gender and number agreement the F-score of the system is calculated that is explained in section 4.3. The results are shown in table 4.2.

The correctness of the system is measured by the language experts. From the table 4.2.

Table 4.2: Results based on gender and number agreement.

Data set	No of sentences	Total words	No of pronouns	Resolved pronoun	Correctly resolved pronoun	F- score
News	12	137	7	7	6	86%
Children story	34	445	28	22	18	72%

It is noticed that pronouns are ambiguous to person, number and gender features While some pronoun can refer to both male and female. These all features affect the performance and F score. . The F score of the news article contains is 86% and The F score is 72 % for story. The F-score of the overall system based on gender and number agreement is 79%. It is examined that the F-score varies with the structure of sentences. The datasets are complex and narrative style and Hindi is free order. So it affects the combine rules of gender and number agreement .It is also observed that sometimes, demonstrative pronouns (वह , यह), Relative pronouns (जिसमें) ,second person pronouns are not resolve correctly and It is observed that certain pronouns refer to both male and female which results the referring to wrong antecedent.

4.4.2 <u>Results Based on Gender , Number Agreement, Co-reference and Animistic knowledge Using Hybrid Approach</u>

Evaluating the performance of the anaphora we have used three different data set described in chapter 4, section 4.1. Total data consists of 4 news articles with 45 sentences, 581 words, 30 pronouns in data set 1. In data set 2 data consists of children story with 34 sentences, 444 words and 28 pronouns. In data set 3 data consists of 2 bibliographies with 71 sentences, 824 words and 67 pronouns. The sentences are complex and have all types of pronouns like first, second, third, reflexive etc. The proposed system recognized the pronouns according to hybrid approach as described in chapter 3 are shown in table 4.3.

Table 4.3: Dataset by Category

Data set	**No of sentences**	**Total of words**	**No of pronouns**	**Resolved**	**Correctly**

				Pronouns by system	**resolved pronouns**
News	45	581	30	37	23
Story	34	444	28	33	25
Bibliography	71	824	67	62	52

We have computed the result by using various metrics are shown in table 4.3. The minimum link between the GOLD and SYS is evaluated by using the MUC metric. It gave 75.85% result on news article, 91.52% on story and 86.15% on bibliography. It gave higher results because singletons are considered in it. Singletons are ignored for false positive while consider for false negative in the B3 evaluation and it gave 65.67% result on the news, 90.32% on the story and 80.99% on the bibliography. To compute the one to one mapping between the entity in the SYS and GOLD the CEAF metric is used which gave 75.45% result on news, 85.89% on story and 72.75%on bibliography. The F- score is also computed and it gave 68.62% on news, 81.96% on story and 80.61% on bibliography.

Table 4.4: Results of hybrid approach on data sets

	MUC	**B3**	**CEAF**	**F-score**
News	75.85%	65.67%	75.45%	68.62%
Story	91.52%	90.32%	85.89%	81.96%
Bibliography	86.15%	80.99%	72.75%	80.61%

The results are represented in graph in figure 4.1. As we seen the MUC of the news articles is less than other domain because the news article contains complex pronouns and less no of singletons. The B3, CEAF, F-score of the story is high as comparative to other domain. Generally in the previous work other researches computed the result on the bases of F-score metric. Other researchers have 61-65% F-score on the news domain and our system gave 68.62%. They have 60-64% and 76-81% F-score on story and bibliography domain and our system gave 81.96% and 80.61%. The overall performance is calculated in table 4.5.

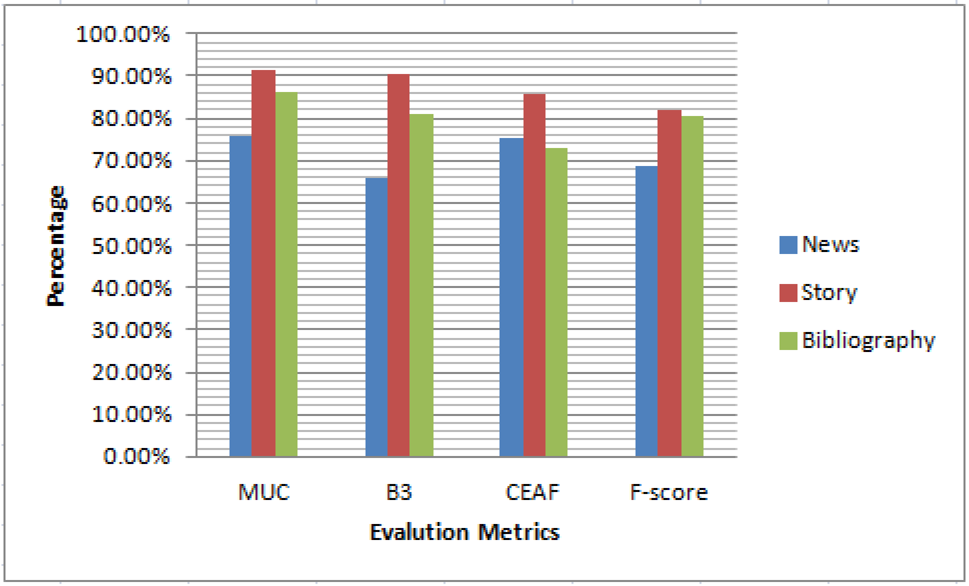


Figure 4.1: Result of evaluation metrics on datasets

In this we can see that the story domain has higher percentage rather than other because it is a straightforward narrative style with extremely low sentence structure complexity. It is observed that success rate of solving the pronoun varies with the structure of sentences. Hindi has no proper structure so the success rate depends on the style of writing. The different article domain has different way of writing that affects the performance of the system.

The overall performance of the system according to the different metrics is MUC gave 84.50, B3 is 78.9%, CEAF is 78.03% and F-score is 77.06%. As compare to previous work the overall performance is calculated by using F-score that is 60 to70% and our system gave 77.06 % which means it is higher accurate than others system.

Table 4.5: Average result of overall system

Evaluation metrics	**Overall performance of system**
MUC	84.50%
B3	78.9%
CEAF	78.03%
F-score	77.06%

We have computed the overall system result with other evaluation metrics. MUC gave the higher results rather than other metrics as shown in figure 4.2 and B3, F-score and CEAF scores are usually lower than MUC on datasets Because in B#. F-score and CEAF singletons are annotated because a great percentage of the score is simply due to the resolution of singletons.

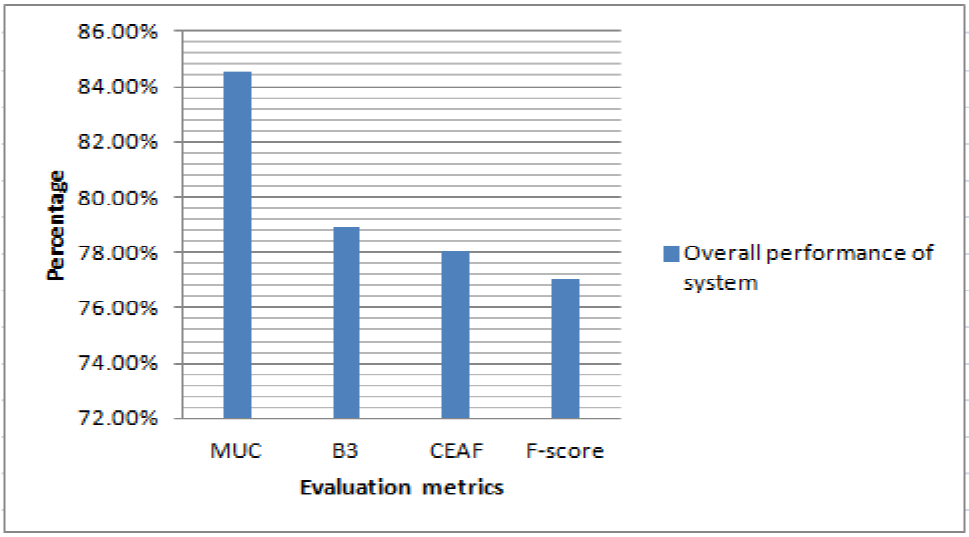


Figure 4.2: overall performance of the system

4.4.3 Result of the agreement

The agreement on behave of judges the accuracy of the system is calculated with the evaluation metric kappa that is designed for categorical judgments and corrects a simple agreement rate for the rate of chance agreement. We conducted the experiment over 3 data sets annotations by 2 raters. Annotators were asked to assign categories like relevant, non relevant etc as stated, according to the type of entity it refers to. The result is shown in table 4.6.

Table 4.6 : Kappa statics for Dataset

Data set	**P(A)**	**P(E)**	**Kappa**
News	0.9729	0.7662	0.6532
Story	0.9090	0.6836	0.7124
Bibliography	0.9354	0.7997	0.6775

The overall kappa for our system is 0.681. According to the table 4.1 the strength of our agreement is substantial.

Chapter 5 : Conclusion and Future Work

Finally, to summarize, this thesis analyzes the benefaction of various researchers who worked on various research issues. Through literature survey we were able to identify various research gaps in anaphora resolution like recency factor, Animistic knowledge, Gender, Number agreement, NER, Pronoun resolution etc.. It can be seen that the problem of anaphora resolution is challenging but not uncontrollable. The thesis describes the definition, type and challenge of Anaphora resolution, various approaches like Rule based, Corpus based, Knowledge poor, Discourse and Hybrid. From last few years anaphor resolution has gain a large attention; a large amount of work has demonstrated which showed good results but in Hindi language a less amount of work has been done. On the basis of related work (in Hindi) , Firstly the thesis proposes an algorithm by using rule based approach to resolve pronouns based on gender and number agreement which gave 79% of accuracy in terms of F-Score.

Secondly, it proposes a hybrid approach which is a combination of Rule based and Learning based to resolve gender, number, co-reference and animistic knowledge. The overall system performance in terms of MUC, B3, CEAF and F-score was observed to be 84.50, 78.9%, 78.03% and 77.06% respectively. The proposed system produced better results than other algorithms. Though the system performance is dependent on the structure of the sentences as Hindi language does not have any standard structure.

5.1 Contribution

The contribution of this thesis is explained below:

- We have studied the previous work in English as well as Hindi language to drill out the research gaps in anaphora resolution.
- We have tried to make a better algorithm which can work on the four issues of the anaphora resolution that are gender, number agreement, recency and animistic knowledge for Indian language.
- The researchers used only one metric that is F- score to measure the performance. But we computed our result by using other metrics that are, BLANC, MUC, B3, CEAF and these metric exhibits significantly different behaviors from each other and globally accepted.
- We have used dependency structure and training tools for resolving the anaphora.

However, apart from gender and number, coreference resolution, recency, animistic there are many issues like intrasentential, intersentential, entity and event anaphora etc also play important role in anaphora resolution. In the future we will try to include all constraint sources to further increase the performance. And none of them are able to cover all issues of anaphora resolution. We can wish to have better results with a time and anaphora resolution approaches for Indian languages.

Publications

- Ashima Kukkar, Rajni Mohana,” Anaphora Resolution in Hindi: Issues and Directions”, paper accepted for “Indian journal of science and Technology (SCOPUS INDEXED JOURNAL)”.
- Ashima Kukkar, Rajni Mohana,” Improving Anaphora Resolution by Resolving Gender and Number Agreement in Hindi Language Using Rule Based Approach”, paper accepted for “Indian journal of science and Technology (SCOPUS INDEXED JOURNAL)”.

Appendix

Data set	Judge 1 say	Judge 1 say	Judge 1 say	Judge 1 say	Total no of

	no but judge 2 say yes	yes and judge 2 say yes	yes , judge 2 say no	no and judge 2 say no	pronouns
News	1	29	5	2	37
Story	1	25	2	5	33
Bibliography	2	53	2	5	62

Judge 1 ____________________

Judge 2 ____________________

References

1. J. L. Vicedo and A. Ferr´andez. Importance of pronominal anaphora resolution in question answering systems.In Proceedings of the 38th Annual Meeting on Association

for Computational Linguistics, page, 555–562. Association for Computational Linguistics, 2000.

2. Wikipedia. Anaphora (linguistics)— Wikipedia, the free encyclopedia, 2014. [Online; accessed 17-January-2014].
3. P. Lakhmani, S. Singh, P. Mathur "Gazetteer Method for Resolving Pronominal Anaphora in Hindi Language", in proceeding of International Journal of Advances in Computer Science and Technology, Volume 3, No.3, March 2014.
4. K. Mehla, Karambir and A. Jangra"Event Anaphora Resolution in Natural Language Processing for Hindi text", in proceeding of IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 1, January 2015
5. T. Lal, P.Kamlesh D.Pardeep "Anaphora Resolution in Hindi: Issues and Challenges", in proceeding of I*nternational Journal of Computer Applications* 42(18):7-13,March 2012
6. Wikipedia. Automatic summarization— Wikipedia, the free encyclopedia, 2014. [Online; accessed 17-January-2014].
7. J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Jeek. Two uses of anaphora resolution in summarization.Inf. Process. Manage., 43(6):1663–1680, Nov. 2007.
8. L. HIRSCHMAN and R. GAIZAUSKAS. Natural language question answering: the view from here. Natural Language Engineering, 7:275–300, 12 2001.
9. R. Mitkov. Introduction: Special issue on anaphora resolution in machine translation and multilingual nlp.Machine translation, 14(3):159–161, 1999.
10. A. Davison. Lexical anaphors and pronouns in hindi. In Lexical Anaphors and Pronouns in Selected South Asian Languages: A Principled Typology, 2003.
11. Kamlesh Dutta , Saroj Kaushik "Anaphor Resolution Approaches :, *Web Journal of Formal Computation and Cognitive Linguistics*, vol.10, pp. 71-76, Jan. 2008.
12. Hobbs, Jerry, "Resolving pronoun references",*Lingua*, vol.44, pp. 311-338, Jan.1978.
13. Carter, David M "A shallow processing approach to anaphor resolution", PhD thesis, Univ. of Cambridge, 1987.
14. Carbonell, J.G. and Brown, R.D., 1988, August. Anaphora resolution: a multi-strategy approach. In *Proceedings of the 12th conference on Computational linguistics-Volume 1* (pp. 96-101). Association for Computational Linguistics.
15. Dagan, I. and Itai, A., 1990, August. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th conference on Computational linguistics-Volume 3* (pp. 330-332). Association for Computational Linguistics.

16. Mitkov, R., 1998, August. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2* (pp. 869-875). Association for Computational Linguistics.
17. Mitkov, R., 1996. Anaphora resolution: a combination of linguistic and statistical approaches. In *Proceedings of the Discourse Anaphora And Resolution Colloquium, DAARC96*.
18. Nasukawa, T., 1994, August. Robust method of pronoun resolution using full-text information. In *Proceedings of the 15th conference on Computational linguistics-Volume 2* (pp. 1157-1163). Association for Computational Linguistics.
19. Connolly, D., Burger, J.D. and Day, D.S., 1997. A machine learning approach to anaphoric reference. In *New Methods in Language Processing* (pp. 133-144).
20. Boguraev, Branimir, Christopher Kennedy,"Salience based content characterization of documents", *ACL'97/EACL'97 workshop on Intelligent scalable text summarization*, 3-9, Madrid, Spain, 1997
21. Christopher Kennedy, Branimir Boguraev,"Anaphora for Everyone: Pronominal Anaphora Resolution without parser", in Proc. *16th International Conference on Computational Linguistics),* Kopenhagen ,Vol.1, august, 1996, pp.113-118.
22. Ruslan Mitkov, "Towards more consistent and Comprehensive evaluation in anaphor resolution. In Proc. *LREC'2000, Athens,* Greece, 2000 , pp.1309-1314.
23. Roland S, "Resolving anaphor References on Deficient Syntactic descriptions", *ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 30-37,Madrid ,Spain, 1997.
24. Roland Stuckardt, "Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm",*Computational Linguistics* ,Vol. 27, pp.445-452, Dec 2001.
25. Baldwin, Breck, "CogNIAC: high precision core- ference with limited knowledge and linguistic resources" *ACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 38-45, Madrid, Spain, 1997.
26. Kameyama, Megumi, "Recognizing referential links: an information extraction perspective",*ACL'97/EACL'97 workshop on Intelligent scalable text summarization*, 3-9, Madrid, Spain, 1997.
27. Brennan, S, Mfriedman, C Pollard, "A centering approach to pronouns", in Proc. *25th Annual Meeting of the ACL*,USA, 1987.

28. Joel R. Tetreault, “A Corpus-Based Evaluation of Centering and Pronoun Resolution" *Computational Linguistics* ,Vol. 27, Issue 4 - Special Issue on Computational Anaphor Resolution, pp. 507- 520, Dec.2001

29. Kolachina, P., Kolachina, S., Singh, A.K., Husain, S., Naidu, V., Sangal, R. and Bharati, A., 2010, May. Grammar Extraction from Treebanks for Hindi and Telugu. In *LREC*.

30. Deepti Chopra1 Dr. G.N. Purohit2”Handling Amb- iguities And Unknown Words In Named Entity Recognition Using Anaphora Resolution”, *International Journal on Computational Sciences & Applications IJCSA*, vol.3, pp.456-463, Oct. 2013.

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

32. Wikipedia. Part-of-speech tagging— Wikipedia, the free encyclopedia, 2014. [Online; accessed 17-January-2014].

33. A. Bharati, D. M. Sharma, L. Bai, and R. Sangal. Anncorra : Annotating corpora guidelines for pos and chunk annotation for indian languages. Technical report, LTRC, IIIT-Hyderabad, 2006.52 N. Chomsky. Syntactic structures. Walter de Gruyter, 2002.

34. N. Chomsky. Syntactic structures. Walter de Gruyter, 2002.

35. Sharma, D.M., Sangal, R., Bai, L., Begam, R. and Ramakrishnamacharyulu, K.V., 2007. AnnCorra: TreeBanks for Indian Languages. *Annotation Guidelines (manuscript), IIIT, Hyderabad, India.*

36. R. Begum, S. Husain, A. Dhwaj, D. M. Sharma, L. Bai, and R. Sangal. Dependency annotation scheme for indian languages. In Proceedings of IJCNLP, 2008

37. A. Bharati, V. Chaitanya, R. Sangal, and K. Ramakrishnamacharyulu. Natural language processing: a Paninian perspective. Prentice-Hall of India New Delhi, 1995.

38. A. Bharati, R. Sangal, and D. M. Sharma. SSF: Shakti Standard Format Guide. LTRC, IIIT-Hyderabad,India, 2007.

39. A. Bharati, S. Husain, B. Ambati, S. Jain, D. Sharma, and R. Sangal. Two semantic features make all the difference in parsing accuracy. Proc. of ICON, 8, 2008.

40. J. Hobbs. Resolving pronoun references. In Readings in natural language processing, pages 339–352.Morgan Kaufmann Publishers Inc., 1986.

41. S. Lappin and H. J. Leass. An algorithm for pronominal anaphora resolution. Computational linguistics,20(4):535–561, 1994.

42. M. A. Walker, A. A. K. Joshi, and E. E. F. Prince. Centering theory in discourse. Oxford University Press,1998.

43. D. Connolly, J. D. Burger, and D. S. Day. A machine learning approach to anaphoric reference. In New Methods in Language Processing, pages 133–144, 1997.

44. Bhargav Uppalapu, Dipti Misra Sharma “Pronoun Resolution For Hindi” in Proc. *DAARC2009,*vol. 5847, April 22, 2009

45. R. Prasad and M. Strube. Discourse salience and pronoun resolution in hindi. U. Penn Working Papers inLinguistics, 6:189–208, 2000.

46. A. Bjrkelund and J. Kuhn. Phrase structures and dependencies for end-to-end coreference resolution. In Proceedings of COLING 2012: Posters, pages 145–154. The COLING 2012 Organizing Committee, 2012.

47. R.M.K. Sinha , A.Jain“Angla Hindi: An English to Hindi Machine-Aided Translation System”*IJET-IJENS* vol. 11 pp. 04 151, 2003.

48. Triveni Lal, Pal,Kamlesh Dutta,Pardeep “Anaphora Resolution in Hindi: Issues and Challenges” *International Journal of Computer Applications*, vol. 4218, pp.7-13, Mar. 2012.

49. Priya Lakhmani1 and Smita Singh2“Anaphora Resolution in Hindi Language”, *International Journal of Information and Computation Technology,.ISSN* , vol. 3, pp. 609-616, 2013

50. Smita Singh, Priya Lakhmani,Dr.Pratistha Mathur and Dr.Sudha Morwal “Anaphora Resolution In HINDI Language Using Gazetteer Method”,*International Journal on Computational Sciences & Applications IJCSA,* vol.4, pp.567-569, Ju. 2014

51. Dutta, K., Prakash, N. and Kaushik, S., 2008. Resolving pronominal anaphora in hindi using hobbs algorithm. *Web Journal of Formal Computation and Cognitive Linguistics*, *1*(10), pp.5607-5607.

52. Uppalapu, B. and Sharma, D.M., 2009. Pronoun resolution for hindi. In *7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)* (pp. 123-134).

53. Sobha Lalitha Devi, Vijay Sundar Ram, Pattabhi RK Rao. "A Generic Anaphora Resolution Engine for Indian Languages". in Proc *25th International Conference on Computational Linguistics,* Coling , 2014, pp.67-84.

54. Dakwale,Vandan Mujadia,Dipti M Sharma “A Hybrid Approach for Anaphora Resolution in Hindi Praveen” in Proc *6th International Joint Conference on Natural Language Processing, IJCNLP* , Nagoya, Japan, Oct. 14-18, 2013, pp.80-86.

55. Kamlesh Duttaa, Saroj Kaushikb, Nupur Prakash “Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items”, *The Prague Bulletin of Mathematical Linguistics* , vol. 95 , pp. 33–50, Apr. 2011.

56. 58 58 Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.
57. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
58. Chen, C. and Ng, V., 2013. Linguistically Aware Coreference Evaluation Metrics. In *IJCNLP* (pp. 1366-1374).
59. Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P. and Vilain, M., 1995, November. MITRE: description of the Alembic system used for MUC-6. In *Proceedings of the 6th conference on Message understanding*(pp. 141-155). Association for Computational Linguistics.
60. Recasens, M. and Hovy, E., 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, *17*(04), pp.485-510.
61. Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LRECWorkshop on Linguistic Coreference*, page563–566.
62. Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human LanguageTechnology Conference and Conference on Empirical Methods in Natural Language Processing*,pages 25-
63. Kaur, Sukhnandan, and Rajni Mohana. "A roadmap of sentiment analysis and its research directions." International Journal of Knowledge and Learning 10, no. 3 (2015): 296-323
64. J. Fleiss. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378, 1971.