I.J. Intelligent Systems and Applications, **2015, 02, 56-64**
Published Online January 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2015.02.08

Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text

Kalyani P. Kamune
RKNEC, Department of Computer Science, Nagpur, 440013, India
Email: kalyanikamune24@gmail.com

Avinash Agrawal
RKNEC, Department of Computer Science, Nagpur, 440013, India
Email: avinashjagrawal@gmail.com

***Abstract*—** One of the challenges in natural language understanding is to determine which entities to be referred in the discourse and how they relate to each other. Anaphora resolution needs to be addressed in almost every application dealing with natural language such as language understanding and processing, dialogue system, system for machine translation, discourse modeling, information extraction. This paper represents a system that uses the combination of constraint-based and preferences-based architectures; each uses a different source of knowledge and proves effective on computational and theoretical basis, instead of using a monolithic architecture for anaphora resolution. This system identifies both inter-sentential and intra-sentential antecedents of "Third person pronoun anaphors" and "Pleonastic it". This system uses Charniak Parser (parser05Aug16) as an associated tool, and it relays on the output generated by it. Salience measures derived from parse tree are used in order to find out accurate antecedents from the list of all potential antecedents. We have tested the system extensively on 'Reuters Newspaper corpus' and efficiency of the system is found to be 81.9%.

***Index Terms*—**Natural Language processing, Anaphora resolution, Discourse, Pronominal Resolution, Co-reference, Discourse Modeling, Artificial Intelligence

I. Introduction

Resolution of anaphoric reference is one of the most challenging tasks in the field of natural language processing. It is extremely difficult to give a complete, plausible and computable description of resolution process as we ourselves deal with it only subconsciously and are largely unaware of the particularities. The task of anaphora resolution is even frequently considered to be AI-complete. Anaphora accounts for the cohesion in the text and is active study in formal and computational linguistics alike. Identifying correct anaphora plays a vital role in Natural Language Processing. Automatic resolution of anaphors is crucial task in the understanding of natural language by computers. Understanding of natural language is difficult for computers because natural languages are inherently ambiguous. On the other hand, human beings can easily manage to pick out the intended meaning from the set of possible interpretations unlike computers due to their limited knowledge and inability to get their bearings in complex contextual situations.

Ambiguity can be presented at different level. It can be presented at lexical level where one word may have more than one meaning (e.g. bank, chair, files). It can also be presented at syntactical level when more than one structural analysis is possible. Ambiguity can also be presented at semantic level or pragmatic level. The automatic resolution of ambiguity requires a huge amount of linguistic and extra-linguistic knowledge as well as inferring and learning capabilities, and is therefore realistic only in restricted domains.

A. Basic Notions and terminologies

Cohesion occurs where the interpretation of the some element in the discourse is dependent on that of another and involves the use of abbreviated or alternative linguistics forms which can be recognized and understood by the hearer or the reader .This refers to or replaces previously mentioned items in the spoken or written text.

e.g. "Sita is a teacher. Her dream is to visit Paris."

In the above example, it is very normal to observe that second sentence is related to the first sentence and hence we can say that **cohesion** is present. In the second sentence, 'her' refers to 'Sita'. Now, in the above example, if we replaced 'her' by 'him', or the whole sentence is replaced by some another isolated sentence, cohesion does not occur any more as the interpretation of second sentence is no longer depends on the first sentence. Discourse features an example of anaphora with the possessive pronoun 'her' referring to the previously mentioned noun phrase 'Sita'. [1]

Anaphora is described as cohesion which points back to some previous item. The pointing back word or the phrase is called an **anaphora** and the entity to which it refers or for which it stands is its **antecedent**. The process of determining antecedent for an anaphora is called **anaphora resolution**. When the anaphora refers to an antecedent and both have the same referent in the real world, they are termed **co-referential**. Various terminologies mentioned above like anaphora, antecedent, anaphora resolution and co-referential can be explained well with the help of example as bellow. [3]

e.g. "The King is not here yet but he is expected to arrive in the next half an hour."

In the above example, the pronoun 'he' is an anaphora, 'the king' is its antecedent and ‘he’ and ‘the king’ are co-referential. Here in this example, antecedent is a noun phrase instead of noun.

Hence, we can say that co-reference is an act of picking out the same referent in the real world. It may be possible that in some examples, a specific anaphora and more than one of the preceding(or following) noun phrases may be co-referential thus forming a co-referential chain of entities which have the same referent. **Co-referential chains** partition discourse entities into equivalence classes.

e.g. ‘Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while traveling on a plane.”

In the above example, Sophia Loren, she (from the first sentence), the actress, her and she (second sentence) are co-referential. Co-referential chain partitions discourse entities into equivalence classes. Hence, in above example, co-referential chain can be singled out: {Sophia Loren, she, the actress, her, she}, {Bono, the U2 singer}, {a thunderstorm}, {a plane}.

B. Types of anaphora

The varieties of anaphora are based on the different types of words which refer back to(or replace) a previously mentioned item like Pronominal anaphora, Pleonastic IT, Lexical Noun-phrase anaphora, Zero anaphora, Noun anaphora, Verb anaphora, Adverb anaphora. Depending on the location of the antecedent, inter-sentential (sentence) anaphora and intra-sentential (discourse) anaphora can be observed. Intra-sentential anaphora arises if the anaphora and its antecedent are present in the same sentence. On the other hand, inter-sentential anaphora is exhibited when antecedent is in a different sentence from the anaphora. Reflexive pronouns are the typical example of inter-sentential anaphora. Possessive pronouns can often be used as intra-sentential anaphora too, and can even be located in the same clause as anaphor. In contrast, personal pronouns and noun phrases acting as intra-sentential anaphoras usually have their antecedents located in the preceding clause of the same complex sentence. The distinction between intra-sentential and inter-sentential anaphora is of practical importance for the design of an anaphora resolution algorithm. Syntax constraints could play a key role in the resolution of intra-sentential anaphors.

The remainder of this paper is organized as follows. In section II, we present the issues in anaphora resolution. In Section III, we represent the related work. Section IV states the hybrid approach for anaphora resolution system. Section V explains the system architecture. Section VI shows the system implementation in detail. Section VII shows the Evaluation environment and the results.

II. SSUES IN ANAPHOR RESOLUTION

Basically, there are two main approaches for resolving anaphora: (1) Traditional Approach which usually depends upon linguistics knowledge, and (2) the Discourse-oriented Approach, here the researcher tries to model complex discourse structure and then uses structures for the process of anaphora resolution. Traditional approaches apply linguistics knowledge, in the way of “Preferences” and “Constraints”, in which systems can be proposed as a technique for combining various information sources. Traditional approaches are mainly works in basic three steps as: (1) Deciding search limit or anaphoric accessibility space, (2) apply various constraints, and then (3) apply preferences. [4]

A. Search limit or Anaphoric Accessibility Space:

This represents a limit for searching all possible candidate antecedents for a particular anaphora. System defines the text segments within which antecedent for a particular anaphora can be found. Finding 'Anaphoric Accessibility Space' is a very crucial phase in the process of anaphora resolution. Because, small search limit results in the exclusion of valid antecedents and too broad search limit results in large candidate lists, which ultimately results in erroneous anaphora resolution. Generally, search limit or 'Anaphoric Accessibility Space' is defined as 'n' previous sentence to anaphora, where value of 'n' varies according to the kind of anaphora. According to Ruslan Mitkov (2008), the ideal anaphora resolution system, value of 'n' is 17 i.e. System check 17 sentences away from the sentence in which anaphora is present. 'Anaphoric Accessibility Space' is predefined by the developer, if any anaphoric word is recognized by the system then list of all possible candidates for antecedent within the predefined search limit is found out for that particular anaphora.[12]

B. Constraints

After getting the list of all possible antecedents, several constraints are applied for removing the incompatible antecedent. Usually, constraints hold certain conditions that must be fulfilled by the candidate antecedent, if it fails, then candidate will not be considered possible antecedents for the anaphor. Various information like syntactical, semantic, morphological, and lexical are used to define various constraints.

C. Preferences

After removing all incompatible candidates for antecedent, if the remaining list of antecedent contains more than one antecedent, then preferences are applied for selecting only one potential antecedent. Preference system must be developed by keeping in mind that only a single candidate must remain at the end of process. And these final candidates given by the preference system will be considered as a potential candidate for that particular anaphora. Various information like syntactical, semantic, morphological, and lexical are used to define preference system.

A text contains linguistic data in many forms such as syntactic parallelism, antecedent proximity, gender and number agreement, lexical repetition or c-command restrictions which plays a vital role in the anaphora resolution process. Various methodologies such as statistical and probabilistic models, Knowledge-poor solutions, using corpus-driven methodologies are preferred for resolving anaphora. Opposite to the pure statistical model, some strategic approaches have also been proposed for tracking the antecedent, which can be formalized in terms of rules based on 'preferences' and 'constraints'. Such strategic approach usually combines of 'constraints' and 'preferences'.

The working of anaphora resolution system relies on the set of various anaphora resolution factors. These factors can be either “eliminating” i.e. it does not count some candidates from the list of all possible candidates or “preferential” i.e. it gives more preference to some candidates than remaining candidates. Partition of anaphora resolution factors into preferences and constraints is responsible for the preferences-based and constraints-based architecture in anaphora resolution. Instead of using single preferences-based or constraints-based architecture, in our system we have used the combination of preferences-based and constraints-based architecture, in order to get the more efficient results. We have used this architecture for our system because study shows that anaphora resolution systems based on constraints and preferences can give a successful result when applied to non-dialogue texts.

III. Related Work

The process of anaphora resolution system traditionally relays on the syntactic, semantic, or pragmatic knowledge in order to identify the antecedent of an anaphor. In this research domain, the first syntax-oriented method proposed was Hobbs’ algorithm in 1976. Hobbs’ algorithm checks number and gender agreement between candidates’ antecedent and a specified anaphora from the outcomes of the syntactic tree.

A statistical approach was introduced by Dagan and Itai in 1990, in which an automatic scheme for collecting statistics on co-occurrence patterns in a large corpus is presented. System uses the corpus information in order to disambiguate pronouns. System uses the semantic constraints in order to disambiguate anaphora references and syntactic ambiguities. It uses the statistical feature of the co-occurrence patterns obtained from the corpus in order to find out the antecedent. In the co-occurrence patterns, the antecedent candidates having highest frequency are selected to match the anaphor.[5]

In 1994, Lappin and Leass proposed RAP (Resolution of Anaphora Procedure) algorithm, which is applied to the syntactic representations generated by McCord's Slot Grammar parser, and relies on salience measures derived from syntactic structure. Working of RAP algorithm relays only on the syntactic representations generated by parser, it does not use any semantic information or real world knowledge in choose accurate candidate among the list of all potential candidate antecedents. An intra-sentential syntactic filter is used for removing anaphoric dependence of a pronoun on an N P on syntactic grounds. A morphological filter is used for removing anaphoric dependence of a pronoun on an NP due to non-agreement of person, number, or gender features. A procedure for identifying pleonastic pronouns is also proposed. An anaphor binding algorithm is sued for identifying lexical anaphor. It also assigns salience weight to find out the final antecedent from the list of all possible antecedents. [11]

Baldwin in 1997 represents a high precision pronoun resolution system. System resolves pronouns only when it satisfies very high confidence rules. Nature of the systems is largely domain independent and reflects processing strategies used by humans for general language comprehension. The system assumes that there is a sub-class of anaphora that does not require general purpose reasoning. The system requires information like sentence detection, part-of-speech tagging, simple noun phrase recognition, basic semantic category information like, gender, number, and in one configuration, partial parse trees. In circumstances of ambiguity, system will not resolve a pronoun. [2]

A robust and knowledge-poor approach for resolving anaphora in technical manuals is proposed by Mitkov in 1998. The text of technical manuals are pre-processed by a part-of-speech tagger and then allowed to check against agreement and for a number of antecedent indicators. Each antecedent indicator assigns score to the candidates’ noun phrase and the candidate with the highest score is selected as the antecedent. It can also be applied to the various languages like English, Polish, and Arabic.[8]

By Denber in 1998, the anaphora resolution is achieved by using WordNet ontology and heuristic rules. An algorithm called Anaphora Matcher (AM) identifies both intra-sentential and inter-sentential antecedents of anaphors. By using the hierarchical relation of nouns and verbs in the surrounding context, it founds the information about animacy. They use the anaphora accessibility space of 2 sentences. [6]

In 1999, Claire Cardie and Kiri Wagstaff proposed unsupervised algorithm. They treated the co-reference resolution as a clustering task. It provides a mechanism for coordinating the application of context-independent and context-dependent constraints and preferences for accurate partitioning of noun phrases into co-reference equivalence classes. [4]

In 2001, Mitkov represented that the comparative evaluation for resolving anaphora has to be performed using the same pre-processing tools and on the same set of data. They proposed an evaluation environment for comparing anaphora resolution algorithms which is illustrated by presenting the results of the comparative evaluation on the basis of several evaluation measures. Evaluation workbench for anaphora resolution proposed by them alleviates a long-standing weakness in the field of anaphora resolution: the inability to consistently and fairly compare anaphora resolution algorithms due to the difference of evaluation data used, and also because of the diversity of pre-processing tools used by each system. [9]

In 2002, Mitkov referred the system as the MARS which operates in full automatic mode. The system presented a new, advanced and completely revamped version of Mitkov's knowledge-poor approach for anaphora resolution. MARS include three new indicators like Boost Pronoun, Syntactic Parallelism and Frequent Candidates. [10]

IV. Hybrid Approach for Pronominal Anaphora Resolution

If we resolve anaphora correctly, it significantly increases the performance of the downstream Natural Language Processing applications. Hence, to address this problem of resolving anaphora correctly, we have implemented a Java-based system which uses a hybrid approach for resolving anaphora. A system used for identifying both inter-sentential and intra-sentential antecedents of third person pronouns in their nominative, accusative or possessive case and pleonastic anaphora. In our system, we have consider the search limit/ Anaphoric Accessibility Space of 3 sentences, hence for any anaphora, system will find all the potential antecedents from the 3 sentences preceding the sentence in which anaphora is present, including the sentence in which anaphora is present. System uses Charniak parser (parser05Aug16) as an associated tool, and it relays on the output generated by it.

Instead of using a single monolithic architecture, system uses the hybrid approach which combines constraint-based and preferences-based architectures System read text file as an input and gives it to the sentence splitter. Sentence Splitter used by the system as an associated tool to split the sentences and put the tags like <S>and</S> before and after each sentence respectively, according to the requirement of the parser. The output generated by the sentence splitter is then given to the parser05aug16 for further processing. A Syntactic representation is generated by the parser called as parse tree. This syntactic representation created by the parser plays a vital role for the further processing. Next step is to create the list of anaphora and antecedent. Now, each anaphora form a pair with all the potential antecedents comes in its Anaphoric Accessibility Space i.e. Within 3 sentence from the sentence in which anaphora is presented. For each pronouns and noun phrase in each pair find agreement features (Number, People and Gender). Each created pair of anaphora and noun phrases is then checks for the agreement feature (Person, Gender, Number) in agreement filter. If the given pair fulfills the agreement feathers then allows passing for the further processing, else pair is discarded by the system. The resulted pairs are then filtered through further filtering process to get the correct antecedent for the anaphora.

All the information required by the system is not used generated by the parser, system have to extract certain required information from the output generated by the parser. Next step is to derive required salience measures from parse tree, which is used for the further processing. Apply "Pleonastic Pronoun Filter" to find pleonastic pronouns (It will take "List of Anaphors" as an input). Apply 'Personal Pronoun Filter' for resolving 'Third person Pronouns', which will take list of noun phrases and pronoun as input. Potential Candidates for antecedent are ranked by their "salience weights" and the top one is proposed as the accurate antecedent. Generate output as "Co-referential Pair".

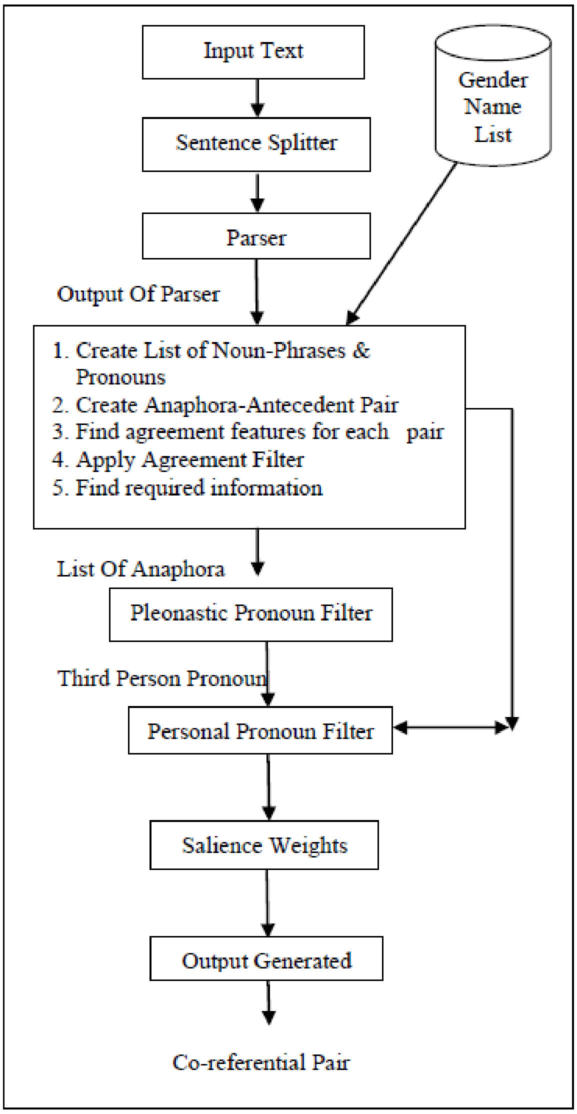
Input Text
Gender Name List
Sentence Splitter
Parser
Output Of Parser
1. Create List of Noun-Phrases & Pronouns
2. Create Anaphora-Antecedent Pair
3. Find agreement features for each pair
4. Apply Agreement Filter
5. Find required information
List Of Anaphora
Pleonastic Pronoun Filter
Third Person Pronoun
Personal Pronoun Filter
Salience Weights
Output Generated
Co-referential Pair

Fig. 1. System Architecture

V. System Architecture

Instead of using a single monolithic architecture, system uses the hybrid approach which combines constraint-based and preferences-based architectures, as shown in the Fig. 1 System takes input in the form of text files and assigns control to the sentence splitter. Sentence splitter spits the sentences and assigns tags in the beginning and end of each sentence as per required by the parser. Output of the sentence splitter is then given to the parser. Parser used by the system is Charniak's Parser (parser05Aug16) which tags the text and generates parse tree. From the syntactic structural generated by the parser, system generates two lists of anaphora and noun phrases. From the list of pronoun and noun phrases, all possible pairs of anaphora and antecedents are generated. Each pair is then filtered through agreement filter which checks for compatibility of each pair on the basis of agreement features. All the information required by the system is not generated by the parser; hence remaining required information is evaluated by the system for the further processing. Next step is to apply pleonastic pronoun filter which will take list of anaphora as an input. After applying pleonastic filter, personal pronoun filter is applied; it considers the list of anaphora and list of noun phrase created by the system. There is a possibility of having more than one potential candidate antecedent for a particular anaphora. So, from the list of all potential candidate antecedents, final antecedent is chosen with the help of salience weight. The architecture of "Pronominal Anaphora Resolution" is given in Fig. 1 in the pictorial form bellow.

VI. System Implementation

A. Sentence Spitter

We take input text and apply Sentence Splitter in order to generate output expected by the Charniak parser. This we use in our system in order to generate parse tree. As the Charnaik Parser expects sentence boundaries to be marked, Sentence Splitter splits sentences as well as apply tags like <S> and </S>, before and after the sentence respectively. We are providing a rule based sentence splitter that efficiently handles input text. It checks text for sentence-ending punctuations like period (.), question mark (?), exclamation mark (!), quotation mark ("") and finally deciding whether to split the sentence there.

Consider bellow example in order to understand the working of sentence splitter.

"As reported, Forrest Gold owns two mines in Western. Australia produces combined 37,000 ounces of gold a year. It also owns an undeveloped gold project."

The output generated by the sentence splitter while considering above example can be given as follows:

<S>As reported, Forrest Gold owns two mines in Western Australia produces combined 37,000 ounces of gold a year. </S>

<S> It also owns an undeveloped gold project. </S>

B. Parser

System uses the publicly available "knowledge rich" Charniak parser (**parser05Aug16**), as input. A parser takes output generated by the 'sentence splitter' as an input and builds a data structure– often some kind of parse tree– giving a structural representation of the input, and checks for correct syntax in the process. All the information required by the system is not given by the parser; system recovers them by using structure information of the verb/noun phrases. Sample output generated by it is shown in Fig. 2.

C. List Of Anaphora and Antecedents

With the help of "Tagged Text" given as an output by the parser, we create two lists as follows:-

- A list of all noun phrases in the input text and
- A list of resolvable anaphors.

```
Input: <s> (``She'll work at the company.") </s>
Output:
        (S1 (PRN (-LRB- -LRB-)
        (S (`` ``)
        (NP (PRP She))
        (VP (MD 'll)
        (VP (VB work) (PP (IN at) (NP (DT the)
        (NN company)))))
        (. .)
        (" "))
        (-RRB- -RRB-)))
```

Fig. 2. Parser Output

D. Pairing

After getting the list of noun phrases and anaphora from the above step, pairing is done within a small sentence window/ Search Limit/ Anaphoric Accessibility Space. Each anaphor is paired with all noun phrases within a small sentence window. Whereas in a given system we are considering a sentence window of 3 sentences i.e. system only considers noun phrases contained within three sentences preceding the anaphor and those in the sentence where the anaphor resides.

E. Agreement features and Agreement Filter

For each anaphora and noun phrase obtained from the created list of anaphora and noun phrases, we find agreement features like Number, People and Gender. To find the agreement features of pronouns/anaphora is straightforward as they are reflected in the pronouns themselves. But it is complicated to find out Agreement features for other noun phrases.

Number: For singular noun phrases the number feature is set as 'true' and for plural noun phrases set it as 'false'. Inspect the tag of the verb phrase, if a noun phrase is found to be agent of a verb phrase. Otherwise, inspect the tag of the noun phrase. Check if the noun phrase has more than one word, if it contains more than one word the existence of the word 'and' or the tag of the phrase's head either can consider to find whether the phrase's number is singular or plural. This feature remains 'unknown' if all these methods fail.

People: In this section, system finds that whether the noun phrase is first person pronoun, second person pronoun, or third person pronoun. By default people feature is considered to be "third". This feature is set as "first", for a plural noun phrase, if in its accusative or nominative case it contains a first person pronoun. If in the nominative/accusative case of noun phrase, second person pronoun is present, the feature is set as "second". This feature remains 'unknown' if all these methods fail.

Gender: English is not so discriminate but in addition to vast majority of neuter words, number of nouns are masculine or feminine or both and failing to identify the gender of such type of words and hence, it can easily lead to errors in resolving anaphora. An extra knowledge base is required to resolve this constrain. Two Gender name lists (male and female) are used to detect the gender feature of noun phrases. Once the string of the noun phrase is found in one of these lists, its gender feature is set accordingly. Otherwise, it remains “unknown”: there is no default value for gender. Gender Filter requires gender information about anaphora and its antecedent.

Agreement Filter

We get number of pairs of noun and pronoun from the list of noun phrase and pronoun phrase, which are then allowed to filter through ‘Agreement filter’. The agreement features’ compatibility of each and every pair of pronoun and a noun phrase is tested by a *Agreement filter*. It states noun and pronoun pair as non-matching in their agreement features only if at least one agreement feature doesn’t agrees, or else it states them as matching. The value “unknown” is regarded to agree with any value of the feature. The constraint system consists of conditions that must be met, and candidates that do not fulfill these conditions will not be considered possible antecedents for the anaphor.

Consider bellow example:

“Cincinnati Bell Inc said it has started its previously-feather announced 15.75 dlr per share tender offer for all shares of Auxton Computer Enterprises Inc.”

In the above example, 'Cincinnati Bell Inc' and 'it' satisfies the conditions of agreement feature and pass for the father filtering.

F. *Required Information Obtained by Inspecting the Parse Tree Structure*

For the Processing of “Personal Pronoun Filter”, we need gather information from the structure of the generated ‘parse tree’. Various terminologies that will come in the description of “Personal Pronoun Filter are given as follows: (these definitions are given by Lappin and Leass. [11]

Table 1. Syntactical Information Required

1	Phrase P is in the *argument domain* of phrase N	iff	1. Phrase P and phrase N both are arguments of the same head. Ex: Ram*i* seem to want to see him*i*.
2	Adjunct domain of N contains phrase p.	iff	1. Phrase N is an argument of a head H. 2. P is the object of a preposition PREP 3. PREP is an Adjunct of H Ex: She*i* sat near her*i*.
3	P is in the *NP domain* of N	iff	1. N is the determiner of a noun Q and (i)P is an argument of noun Q, or (ii) Preposition PREP has an object P and PREP is an adjunct of noun Q. Ex: Shyam*i*’s poem on him*i* is funny.
4	P is *contained* in a phrase Q	iff	1. N is the determiner of a noun Q and (i)P is an argument of Q, or (ii)Preposition PREP has an object P and PREP is an adjunct of Q. Ex: Shyam*i*’s poem on him*i* is funny.

Where, P: Phrase N: Phrase

The information represented by the above mentioned terminologies can be obtained by inspecting the parse tree as follows:

Table 2. Information fetch from parse tree

1.	NP is in the argument domain of another NP	if	1. One is a child of the following sibling VP of the other, or 2. Two NPs are connected by conjunction and they together form a sibling of a VP. 3. VP is the argument head of both NPs
2.	NP is in the adjunct domain of another NP	if	1. Former is a child of a PP, which is again children of a VP, and 2. Latter is either a sibling or children of the VP. 3. VP is the adjunct head of the former NP
3.	An NP is in the NP domain of another NP	if	1. Former NP is a child of a PP 2. PP has a proceeding sibling NP, children of which include the later NP and a following POS.
4.	NP is contained in a VP NP is contained in another NP NP is considered to be contained in a VP/NP	if if if	1. VP is the NP’s argument head or adjunct head; 1. Former is a child of the latter’s sibling PP 1. It is contained in a phrase Q and Q is contained in the VP/NP.

G. *Pleonastic Pronouns Filter*

In addition to the first and second person pronoun, the pronoun it can often be non-anaphoric. For example, in below sentence, it is not specific enough to be considered as anaphoric:

e. g. “I fear it may rain.”

Non-anaphoric uses of it are also referred to as pleonastic or prop it. Examples of pleonastic ’it’ include non-referential instances. Non- anaphoric uses of ‘it’ are not always a clear cut case and some occurrences of it appear to be less unspecified than others and are therefore a matter of debate in linguistics. The pronoun “it” is commonly used as the pleonastic pronoun in English language. Typically it appears with a *modal adjective* or a *cognitive verb* in its passive participle form. A System uses the list of modal adjective and a cognitive verb in order to find pleonastic pronouns appearing in the predefined syntactic patterns. AAR system performs a pattern matching for each keyword ‘it’ and it is declared as pleonastic if the matching is successful.

H.Personal Pronoun Filter

It is the most wide spread type of anaphora. Pronominal anaphora occurs at the level of personal pronouns, possessive pronouns, reflexive pronouns. The set of anaphoric pronouns consists of all third person personal (he, him, she, her, it, they, them), possessive (his, her, hers, its, their, theirs), reflexive (himself, herself, itself, themselves) pronouns in both singular and plural. Whereas, first and second person pronoun can often be non-anaphoric.

System is used to find out the “Third person pronouns” in their nominative, accusative or possessive case. Information obtained from the Parse Tree Structure plays a vital role for identifying the “Third person pronoun anaphora”.

e.g. “Sumitomo President Koh Komatsu told Reuters he is confident his bank can quickly regain its position.”

In the above example, 'he' and 'his' both refers to the 'Sumitomo President Koh Komatsu' were correctly identified by the personal pronoun filter.

I. Salience weights

After applying agreement filter and personal pronoun filter a, there is a possibility that for a particular anaphora more than one potential candidate antecedent can be present. So, from the list of all potential antecedents which we obtained after applying agreement filter and personal pronoun filter final accurate candidate antecedent is chosen with the help of salience weight proposed by Shalom Lappin and Herbert J. Leass. Salience weight is assign to all potential candidate antecedents, and the one having highest salience weight is chosen finally for the anaphora antecedent pair.

J. Inter-sentential and Intra-sentential Anaphora

Both inter-sentential (sentence) anaphora and intra-sentential (discourse) anaphora is identified by the system. Intra-sentential anaphora arises if the anaphora and its antecedent are present in the same sentence. On the other hand, inter-sentential anaphora is exhibited when antecedent is in a different sentence from the anaphora. Search limit or 'Anaphoric Accessibility Space' is defined as the 'n' previous sentence to anaphora, where value of 'n' varies according to the kind of anaphora. In our system we have consider the 'Anaphoric Accessibility Space' or the value of 'n' to be 3. Hence, if any anaphora is found by the system, it will check for its candidate antecedents in the 3 sentences previous to the sentence in which anaphora is present along with sentence in which anaphora is present.

Consider an example as follows:

“The analysts agreed the bank was aggressive. It has expanded overseas, entered the lucrative securities business and geared up for domestic competition, but they questioned the wisdom of some of those moves.”

In the above example, 'It' refers to the 'Bank' and ' they' refers to the 'analysts' were correctly identified by the system. Above example represented the inter-sentential (sentence) anaphora. Whereas for the intra-sentential (discourse) anaphora consider the example as given bellow:

“Locke said shareholders would be advised as soon as the discussions progressed and recommended that they keep their shares.”

In the above example, 'they' represents the 'shareholders' was correctly identifies by the system. [7]

VII. Evaluation

Testing of System performs manually. For the testing purpose, we have use 'Reuters Newspaper corpus'. In Reuters Newspaper corpus', total 4024 files are present in the test section and total 11,413 files are present in the Training section. In 'Reuters Newspaper corpus', files are present in total 91 various categories like housing, income, jobs, money-supply, livestock, retail, interest, silver, trade and so on in both test and training section. In the 'Reuters Newspaper corpus', total 15,437 files are present distributed over 91 different categories in both test and training section. Files in the 'Reuters Newspaper corpus' are present in TXT format, of minimum size 693 byte and maximum size of 4.3 kb approximately.

Table 3 shows that testing of system is performed on total 120 file of different categories of 'Reuters Newspaper corpus'. In these 120 files, total 442 anaphora-antecedent pairs were found out of which 362 pairs were correctly identified by the system. Hence, we can say that efficiency of the system is 81.9%. 120 files which we have used for the testing purpose are present in TXT format, of minimum size 693 byte and maximum size of 4.3 kb approximately. We can say that the average size of each file is 2.5 kb. On an average, we can say that each file contains approximately 4 pairs of anaphora-antecedent pair.

Table 3. System Performance

Total File considered	Total Number of Anaphora-antecedent Pairs Present	Number of cases that the system resolves correctly	Accuracy %
120	442	362	81.9 %

In total 442 pairs of Anaphora-antecedent, all third person personal (he, him, she, her, it, they, them), possessive (his, her, hers, its, their, theirs), and reflexive (himself, herself, itself, themselves) pronouns in both singular and plural, are tested along with the occurrence of pleonastic it. From the result coming out from the experiment, system correctly founds total 27 pairs of third person personal from 45 pairs , 144 pairs of possessive pronouns from 167 pairs, and 191 occurrence of ‘It’ pronouns from 230 pairs in both singular and plural form as shows in Fig. 3. Hence we can say that system have 60% efficiency to find third personal, 86.22% efficiency to find possessive pronoun and 83.04% efficiency in finding 'It' pronoun

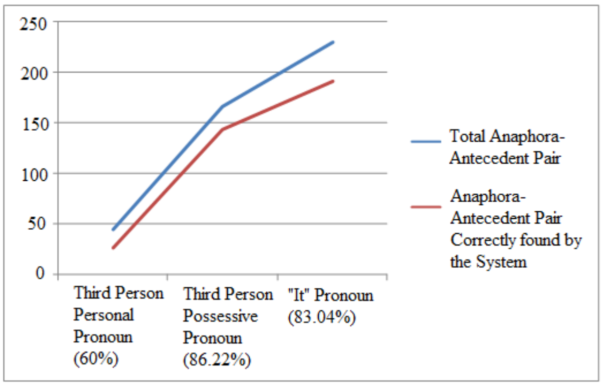

Fig. 3. Performance of System

We can say that system almost founds of third person personal, 144 pairs of possessive and 191 pairs of reflexive pronouns In order to improve the performance of the system powerful personal pronoun and agreement filter plays an important role. The way in which implementation extract the grammatical roles by applying certain hand-crafted rules on the parse tree also affect the overall performance of the system. Use of knowledge rich charnak parser (parser05aug16) in system, also contributes in overall performance of the system.

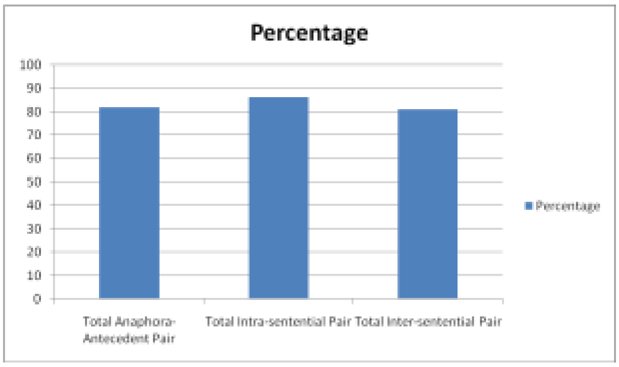

Fig. 4. Efficiency of System in Percentage

Along with the intra-sentential anaphora system also finds inter-sentential anaphora. Out of the total 442 pairs of anaphora and antecedent, total 362 pairs were intra-sentential and total 80 pairs were found to be inter-sentential. Out of total 362 intra-sentential pairs, system founds 293 anaphora-antecedent pair correctly and out of total 80 inter-sentential pair system finds 69 pairs correctly. Hence, efficiency of system to find inter-sentential and intra-sentential anaphora-antecedent pairs is 86.25 and 80.93 respectively, which is shown by Fig. 4 in a pictorial form.

VIII. Conclusion and Future Work

System uses the hybrid methodology which combines both constraint-based and preferences-based architectures, for resolving anaphors which is represented as above. System works efficiently in order to identify inter-sentential and intra-sentential antecedents of "Third person pronoun anaphors" and "Pleonastic it". The System at first defines an anaphoric accessibility space or search limit, then applies various constraints, and finally applies preferences for identifying correct antecedents. The gender features extraction methodology used by the system in the discourses is helpful to the promotion of resolution accuracy. The uses of knowledge rich Charniak parser are helpful to the promotion of the resolution accuracy. The proposed system which uses hybrid approach is able to deal with intra-sentential and inter-sentential anaphora in English text and includes an appropriate treatment of pleonastic pronouns. In contrast to most anaphora resolution approaches, our system operates in fully automatic mode to achieve optimal performance. Along With the growing interest in the field of natural language processing and its applications in various fields, anaphora resolution is worth considering for further language understanding and the discourses modelling.

Acknowledgment

I take this opportunity to express my deepest gratitude and appreciation to all those who have helped me directly or indirectly towards the successful completion of this project. I want to thank God and my parents. I would like to express my sincere appreciation and gratitude to my guide Dr. Avinash Agrawal, Professor, Department of Computer Science & Engineering, RCOEM, Nagpur for his insightful advice, encouragement, guidance, critics, and valuable suggestions throughout the course of my project work. Without his continued support and interest, this thesis would not have been the same as presented here.

References

[1] Aarts Jan, Henk Barkema and Nelleke Oostdijk (1997), "The TOSCA-ICLE Tagset: Tagging Manual", TOSCA Research Group for Corpus Linguistics.

[2] Baldwin, Breck (1997), "CogNIAC: high precision coreference with limited knowledge and linguistic resources", Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution

[3] Cardie, Claire and Kiri Wagstaff (1999), "Noun Phrase Coreference as Clustering", Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[4] Chinatsu Aone and Scott William Bennett, "Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies", *International Workshop on Sharable Natural Language Resources (SNLR)*,2000

[5] Dagan, Ido and Alon Itai (1990), "Automatic processing of large corpora for the resolution of anaphora references", Proceedings of the 13th International Conference on Computational Linguistics (COLING'90), Vol. III, 1-3, Helsinki, Finland.

[6] Denber, Michel (1998), "Automatic resolution of anaphora in English", Technical report, Eastman Kodak Co.

[7] Mitkov and Ruslan, "*Anaphora resolution in Natural Language Processing and Machine Translation". Working* paper. Saarbrücken: IAI, 1995a.

[8] Mitkov, Ruslan, "Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches" *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 14-21. Madrid, Spain, 1997b.

[9] Mitkov, Ruslan and Catalina Barbu (2001), "Evaluation tool for rule-based anaphora resolution methods", Proceedings of ACL'01, Toulouse, 2001.

[10] Mitkov, Ruslan, Richard Evans and Constantin Orasan (2002), "A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method", In Proceedings of CICLing- 2000, Mexico City, Mexico.

[11] Shalom Lappin and Herbert J. Leass ,"An Algorithm for Pronominal Anaphora Resolution", 1994

[12] Ruslan Mitkov, "ANAPHORA RESOLUTION: THE STATE OF THE ART", *International Conference on Mathematical Linguistics,* 2008

Authors' Profiles

Kalyani P. Kamune, born on 24th March 1991 in Nagpur. She received her Bachelor of Engineering in Computer Technology in 2008. She completed her M.Tech from Computer Science in 2014 from Shri Ramdeo baba College of Engineering and Management, Nagpur, India. Her research areas are **Natural** Language Processing, and Artificial Intelligence.

Dr. Avinash Agrawal. He received his Bachelor of Engineering in Computer Science. He completed her M.Tech from Computer Science. Currently he is working as a Professor in Department of Computer Science and Engineering at Shri Ramdeo baba College of Engineering and Management, Nagpur, India. His research areas are **Natural** Language Processing, and Artificial Intelligence.

How to cite this paper: Kalyani P. Kamune, Avinash Agrawal,"Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text", International Journal of Intelligent Systems and Applications (IJISA), vol.7, no.2, pp.56-64, 2015. DOI: 10.5815/ijisa.2015.02.08