# INFORMATION EXTRACTION AND AD HOC ANAPHORA ANALYSIS

**[1]A.SURESH BABU, [2]DR P.PREMCHAND, [3]DR A.GOVARDHAN**

[1]Asst. Professor, Department of Computer Science Engineering, JNTUA, Anantapur
[2]Professor, Department of Computer Science Engineering, Osmania University, Hyderabad
[3]Professor, Department of Computer Science Engineering, JNTUHCEJ, Jagityala

**ABSTRACT**

Anaphora is the prominent linguistic element that defines the contextual coreference. It is very essential to apprehend and reason out large texts for the discovery of summary and forms of knowledge. Methods and Techniques are available for specific contextual elements that frame out as summaries and forms of knowledge. Important information which have keen actionability is sometimes not possible to extract using methods and techniques that are approved from available research. And, this is a task which shall be requisite for text detection and extraction to find out reasonable inferences from a huge corpus of non domain texts. We propose Ad Hoc Anaphora Analysis as the novel method introduced in this work to structure and build the framework that implement for a large non domain text corpus.

**Keywords:** *Linguistic Analysis, Coreference, Anaphora, Ad Hoc Modeling, Uncertainty Models.*

## 1. INTRODUCTION

An increasing demand for Information Extraction (IE) systems and the tendency for improving their performance (precision and recall) call for context and coherence analysis. The coherence can be obtained by different linguistic phenomena among which anaphora plays a crucial role. The first gigantic step is MUC, which was to boost wide research in automatic information extraction and defines the research for the decades of future even. Most of the MUC systems cover very limited subjects and on the other hand, are capable of handling texts belonging to heterogeneous text types and subject domains, and usually use very generic classification schemes, which might be refined, if the information processing task demands a more specific identification of semantic information. Anaphora analysis for information extraction purposes was taken into consideration during the Message Understanding Conferences (MUC); the conferences were dedicated to information extraction systems, and finally formulated as coreference task during information extraction at the 7th MUC. MUC some times called Message Understanding Competition called after its competition character.

This paper concerns anaphora as a language phenomenon, which introduces the connection between pointing back expression anaphor (called referent as well) and antecedent; the present article discuses only anaphors that can take shape of a pronoun – pronominal anaphora. Anaphora is a "*coreference is regarded as type of anaphora where an identity relation, between the anaphor and the antecedent, is preserved*".

Content analysis or text analysis particularly in the documents that are embodied as great text documents is a scholarly methodology by which are text are studied with authorship, authenticity and other text elementaries such as philology, hermeneutics and semiotics. There are many prevailing Text Analytics as core technology offering a wide-range of analysis techniques including statistical and linguistic based approaches. Statistical techniques enable our applications to classify terms at both the sentence and document level. These techniques also enable our end user applications to include statistical analysis like magnitude, mean and cluster analysis. The need of the technology is to automatically apply the "best fit" engine for the analytical problem so that users don't have to understand the analysis approaches but rather focus on the analysis outcomes that result from pre-packaged (but customizable) queries, reports. Proposing a sophisticated engineered framework that has a linguistic approach to enable the applications to advance a step further by providing users with a view of the relationships between people, places and things, thereby allowing users to identify problem root cause or the degree of sentiment.

Presenting statistically text analysis means, summarizing, quantitative analysis of text messages that relies on the scientific methods such as *a priori design*, *consistency*, *reliability*, legality, generalizability, replicability, objectivity and intersubjectivity.

## 2. RELATED WORK

The first effort to define anaphora and create methods for analyzing them was made in the late 70's. Automatic methods from that time based on knowledge engineering – rules which mainly used *syntactic information*. Some of them where supported by late *binding theory*. Following modifications tend to occur *cognitive techniques* and *short memory model* are applied.

The 90's contributed to a better discourse and coherence understanding, which triggered the emergence of algorithms based on *centering theory*. In those times, one made an effort to use corpora for anaphora analysis and machine learning (ML) approach. Even a genetic algorithm for finding salience weights in pronoun resolution system was implemented. In an alternative corpus-based approach a large number of documents is processed for *statistical analysis*.

Recent scientific research tends to apply shallow parsing and rules which operate on empirical *antecedent indicators* independent of language; it achieved almost has attained 90% success rate. There are studies on decision tree for coreferences for deducing definite descriptions and some bridging anaphoras.

A further step in plain text analysis is the distinction between dictionary-based (quantitative) approaches and qualitative approaches. Dictionary-based approaches set up a list of categories derived from the frequency list of words and control the distribution of words and their respective categories over the texts. While methods in quantitative content analysis in this way transform observations of found categories into quantitative statistical data, the qualitative content analysis focuses more on the intentionality and its implications.

### 2.1 Anaphoras in Information Extraction

Understanding the text lexically creates ambiguity, where words may have more than one meaning (e.g. *bank*, *file*, *chair*), but also at the syntactic level when more than one structural analysis is possible (e.g. *Flying planes can be dangerous*, *I saw the man with the telescope*). Furthermore, ambiguity is exhibited at the semantic level (*The rabbit is ready for lunch* – where the rabbit can be interpreted as both agent and patient) or pragmatic level (*Can you open the window?* – where this phrase can act both as a request and as a question, depending on the contextual situation). The anaphor resolution is very important and hence requires a huge amount of linguistic and extralinguistic knowledge as well as inferring and learning capabilities, and is therefore realistic only in restricted domains.

There are many varieties of anaphoras, which should be resolved if one wants to understand the meaning of the text. Apart from the problem of definition and denotation undertaken in the earlier studies, characteristic properties of anaphora are worth emphasizing. Especially the two of them are interesting – a type of non-coreferential relation (*bridges*). The authors enumerate a set of relations among nominal anaphora participants (though the article mentions relation for other types of anaphoras): *set membership*, *part-of* (necessary part, probable part, inducible part), which with relations from ontology (like specialization/generalization), need to be investigated during analysis.

Information extraction is typically faced with the problems driven by recognized entities (during named entity task), the coreference task become important part of processing, extending scope of extraction. The wider scope system has the better results it gets, so capturing information about all mentions of any given random entity (including non-coreferential relations) should be regarded as ad hoc anaphora resolution.

### 2.2 An approach for ad hoc anaphora analysis

As "Text Knowledge Extraction" maps natural language texts onto a formal representation of the facts contained in the texts. Common text knowledge extraction methods show a severe lack of methods for understanding natural language "degree expressions", like "*expensive hard disk drive*" and "*good monitor*", which describe gradable properties like price and quality, respectively. However, without an adequate understanding of such degree expressions it is often impossible to grasp the central meaning of a text. Ad hoc processes only can bid good rating for finding such kind adjectives and transitives from the text. But, most of the instances the text is not

clear and it is very difficult to understand intricate and interesting meanings. Ad hoc processes are tractable when the process heuristics direct non-hypothesis based experimentation. Hypothesis based experimentation drives into a work flow model and follows the component by component process to obtain the expected result. In an ad hoc workflow process, the user decides how a document should be routed for review when the document is selected as candidate for the process. It is very difficult to follow a uniform process of obtaining the resulting anaphora in the document using the ad hoc process. The algorithmic approaches pursue narrow space methods which can derive the results that pre-specific. But the objective and goal of using ad hoc anaphora analysis is to find the anaphora that is hidden from the classical models of information extraction, and find the coreference concepts. An etym is used to define the evolution semantics of a term. But it is very difficult to find algorithmically such terms. Ad hoc anaphora analysis is a challenging issue that can derive the anaphora within the text that contains the coreference related to the etymological significance. The etymological connection of any term can be known only if the word or terms evolution is known. But it is a process of pre-linguistic which related to the evolution of language and vocabulary. As the case of the epical documents the terms are not just colloquial or not present in the usual vocabulary of a vernacular. Such words are extracted with an evolutionary meaning, when known their etymology.

Collecting anaphora is a collocative (*co-locative*) process that includes building of terms (*~etyms*) from reliable, valid, generalized documents. Statistical quantitative analysis determines the frequencies of the word occurrences; however the semantic analysis should be made on the document to find the *sensible terms* (*~etyms*). An ad hoc process should be a statistical and semantic analysis on the document content.

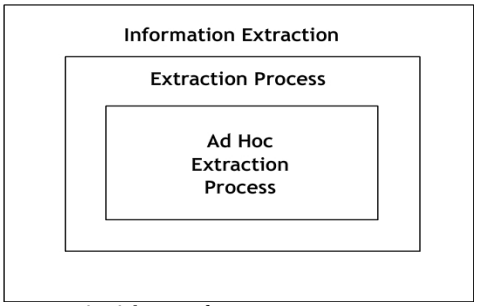


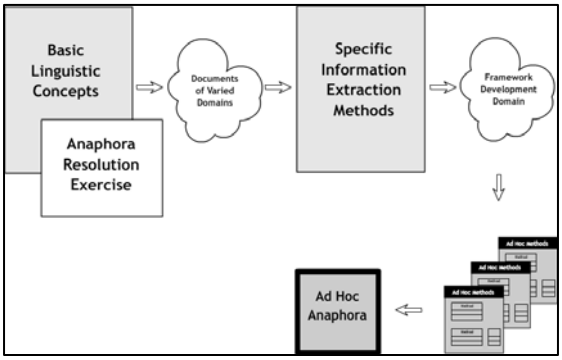

Fig. 1: An Ad Hoc Information Extraction Process.

A collocative process is otherwise collocation process that introduces due methods to derive the following terms (*~etyms*).

### 2.3 Modeling Framework for Ad Hoc Anaphora Analysis

The Meta-Models describe the categorical idea of the models, with respect to fitting location of the model into the place of the IE system, the most chosen models proposed are described, viz., Uncertainty Model, Pragmatic Model, Canonical Model, Matrix Model and other Meta-heuristic models. The matrix model is further subscribed with Matrix and Sub-matrix Model and Cross-Association Model.

Fig. 2: Schematic Overview of Ad Hoc Anaphora Analysis




The Uncertainty Model is a basic reference model for cloud model, with uncertainty between a quality concept which is expressed by natural language and its quantity number expression. If $\boldsymbol{U}$ is a quantity domain expressed with accurate numbers and $\boldsymbol{C}$ is a quality concept in $\boldsymbol{U}$, if the quantity value $\boldsymbol{x} \in \boldsymbol{U}$, and $\boldsymbol{x}$ is a random realization of the quality concept $\boldsymbol{C}$, $\mu(\boldsymbol{x})$ is the membership degree of $\boldsymbol{x}$ to $\boldsymbol{C}$, $\mu(\boldsymbol{x}) \in [0,1]$, it is the random number which has the steady tendency:

$$\mu : U \to [0,1], \forall x \in U, x \to \mu(x)$$

The distribution of $\boldsymbol{x}$ in domain is called cloud model, which is briefly called cloud, each $\boldsymbol{x}$ is called a cloud drop.

The quality anaphora and its contextual significance is development of a right model chosen for ad hoc analysis on text for discovering

anaphora. The uncertainty model describes a best version of ad hoc anaphora analysis.

### 2.4 Other Important Models

Pragmatic model is associated with knowledge representation and its complexity. Sometime knowledge is expressed as a set of rules, of the form: if ***x*** then ***y***. Empirical experiments or surveys, where the aim is to find those factors that distinguish between different outcomes, are an example. Knowledge is also expressed as tables, trees, etc,. In a table form of knowledge the conjunction of values does not predict both that property and its absence, whereas in a tree the branches could be dictated by different responses to a questionnaire. Choosing a predicate of interest will produce predictive rules; if there are no contradictions.

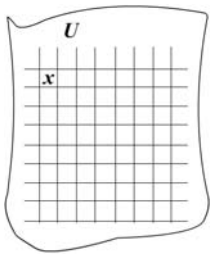



Fig. 3: Uncertainty Model

The *pragmatic model* is selective for its clear resolution abilities when there are contradictions in the consequences; however the consequences are subset of the truth universe. *Canonical Model* of a process that is universally adaptable to any unstructured problems. Problems come as structured and unstructured. Structured problems have well defined results or expectations of the results form is definite, where unstructured problems are unnatural but to be solved by optimizing the problem elements and finding the solutions. *Matrix Model* emphasizes in condensing high dimension data into compact form, which is easy to handle by algorithms that process. *Meta-heuristic models* exist in nature that comap to evolutionary behavior of *phenomenal plurale*.

Either data is converted or available in matrix is condensed to row-references and column-references and the cross-intersections of the row groups and column groups are used to solve the problem, where the optimization and meticulous selection of row and column groups is the key role.

## 3. PROPOSED WORK

### *3.1 The chosen model*

The uncertainty model or the pragmatic model designs the overall ad hoc anaphora analysis. At the kernel of the whole the information extraction from the text is only possible by understanding the text theoretic principles. These principles include more importantly the theory of text structures that separates the text into two segments which processes local as well as global coherence. *Global coherence* expresses the interaction between segments and their composition toward a discourse structure at large. *Local coherence* is responsible for the inter-sentential level and is tightly connected with syntactic, semantic, and positional information from each sentence. Centering is intended to capture local coherence. The principal idea of the centering model is to express fixed constraints as well as "*soft*" rules which guide the reference resolution process with a minimal computational load on the cognitive system of the reader. As it is known from the documents, the information extraction process is in a state of uncertainty with respect to a wide collection of domains. The main data structures of the centering model are a list of *forward-looking centers*, $\boldsymbol{C_f}$ $\boldsymbol{(U_k)}$, and one *backward-looking centers*, $\boldsymbol{C_b(U_k)}$, each for utterance $\boldsymbol{U_k}$. The functional centering model (FCM) is composed of forward-looking centers and backward-looking centers. The former model of the FCM denotes the given information and the later model a theme-hierarchy.

### *3.2 Tracking anaphors*

Identifying the antecedent of an anaphoric trigger (a pronoun, definite DP, etc.) depends on the interaction of many factors: syntactic (e.g. Binding Theory), semantic (e.g. selectional restrictions), and pragmatic (e.g. Centering Theory). Some of these factors, such as selectional restrictions and syntactic binding requirements rule out certain antecedents, while other factors, e.g. topicality, suggest that a certain antecedent should be chosen.

"*India not only requires at least two victories in its remaining three matches to go through, but may also need a bonus point along the way.*" … "*it rained here on Sunday and there could be a cloud cover during Monday's game.*" … "*Even otherwise, there could be a tad more moisture on the surface. An already seamerfreindly pitch could assist the pacemen further.*"

The above three statements are taken from sport extract of "The Hindu", Monday, 16th August 2010. The first line clearly puts that requirement of Indian team, the second line describes about the situation. If the two lines are not known for typical text examination the third line the pacemen would be meaning less. Detecting the topicality, syntactic application and giving inputs to the FCM can determine the players of the match.

Even what appear to be inviolable constraints, such as number agreement, can some times be overruled.

Such examples abound; and they indicate that all anaphora resolution factors, or almost all of them, are best thought of as defaults, which may be overridden. It is therefore attractive to model anaphora resolution as a system of ad hoc defaults.

Most such systems do not encode the constraints explicitly, but rather procedurally, as part of the algorithm. There are, however, string arguments for having a declarative, explicit definition of the constraints. They implement a system of constraints for anaphora resolution.

This paper is not mere identifying the factors or their relative strengths, rather to argue that formalizing all these factors is not enough and additional rule is necessary; hence the formalization of this rule in default logic for ad hoc anaphora analysis.

*3.3 Don't overlook anaphoric possibilities:*

The DOAP is based on Optimality Theoretic system of prioritized defaults for anaphora resolution. We propose default logic for ad hoc anaphora analysis. Consider the above discourse again; the antecedent that is eventually chose, the pacemen, is not suggested by any of the well known factors discuss in the literature: it is neither topical, nor a subject, nor does it have the same syntactic position as the pronoun, etc. This antecedent is simply chosen as a last resort, since the other potential candidate is ruled out for expression. This "last resort" rule must be defined somehow, for, without it, no antecedent would be chosen. Indeed, in the linguistics literature, such a rule has been proposed.[1]

Essentially, this rule says that, when we encounter a trigger, we must try to find an antecedent. If we find an antecedent that is suggested by some rule, so much the better; but even a dispreferred antecedent is better than no antecedent at all. In practice almost all anaphora resolution algorithm obey DOAP, in the sense that they always attempt to find (at least) one antecedent, even if the anaphora is ambiguous. However, if DOAP is not defined explicitly in the object level of the logic, but is left to a metalevel description, it is hard to be clear on, let alone prove, what a system will do when there is no clear choice of antecedent; which, if any, antecedent it will choose, and which inferences it will draw. Hence, formalization of DOAP on a par with all other factors is a desirable goal.

Developing a framework and a software product model is aimed in the work, where the corpus of elements, etyms and the contextual text is preserved and used for ad hoc analyses.

*3.4 Formalization*

The relation between trigger and antecedent is equality, so the problem of anaphora resolution becomes the problem of inferring the necessary equalities from the representation. The ad hoc nature of the linguistic applications requires a broad analysis on the trigger and antecedent. Using the default theory, which uses nonmonotonic formalisms, it is possible to formalize the anaphora resolution in ad hoc linguistic application. These formalisms include a substantial body of theoretical work to be devoted and a number of theorem provers have been used default logic.

One suggestion for representing and reasoning with commonsense knowledge was developed by Ray Reiter [1980] and is known as default theory. The idea is to reason in first order logic but to have available a set of default rules which are used only if an inference can not be obtained within the first order formulation. The general framework of this proposal is depicted below.

The premises consist of two components. The first is a set of first order expressions and is referred to as the prerequisite. These expressions must be proven (in the standard deductive sense) to be true in order for the rule to be applicable. The next set is referred to as the consistency test. These expressions must be consistent with the current database. That is, it must be proven that the negation of the expressions does not follow from the current database. If the rule is proven to be applicable, then the expressions referred to as the consequent are added to the database.

Default theory consists of a pair (D, W) where W is a set of first order formulae and D is a set of default rules of the form;

$$\frac{\alpha(\vec{x}) : M\beta(\vec{x})}{\gamma(\vec{x})}$$

Where; $\alpha(\vec{x})$ is the prerequisite of the default rule, $M\beta(\vec{x})$ is the consistency test of the default rule and $y(\vec{x})$ is the consequent of the default rule.

The rule can be read as:

> "For all individuals $x_1 \ldots x_m$, if $\alpha(\vec{x})$ is believed and if each of $\beta(\vec{x})$ is consistent with our beliefs, then $\gamma(\vec{x})$ may be believed".

Operator *M* refers to consistency with respect to the deductive closure of the set of beliefs.

An Example of Default Rule:

*Had the police taken all the statements they needed from her?*

*He that plants thorns must never expect to gather roses.*

$$\frac{police(x) : M(\exists(y)), statements(x) \wedge client(x)}{statements(x) \wedge client(x)}$$

$$\frac{he(x) : M(\exists(y)), thorns(x) \wedge roses(x)}{thorns(x) \wedge roses(x)}$$

There is no single, distinguished modal logic for describing default reasoning [4]. On the contrary, there exist whole ranges of modal logics, each of which can be used in the embedding as"host" logic. This shows that, in agreement with the intuition, in order to capture default reasoning the most important step is to translate into a nonmonotonic modal system the principle of "negation as failure to prove". Once this is made, then the choice of particular modal axiom schemata is of secondary importance, in fact, there is a large degree of freedom in which of them to choose.

*Implementation:*

The approach described in this paper attempts to develop a typical framework for anaphora analysis. The judgment of anaphora of multi domain texts which is a very irregular in context is the prime job of priority. The construction of framework for ad hoc anaphora analysis is practical with the convergence of corpus (*discourse and theories*) and default logic, statistics, probability and distribution. A rich collection of premises that belong to the facts expressed in different syntactically and semantically relevant contexts are the base for the implementation of default logic to determine the anaphora. The irregular property of the anaphora in the analysis process can be achieved by building a huge corpus of premise facts. A corpus of etyms is for the anaphoricity of words. The efficient search algorithms and parsers have to be employed to detect the skeletal structure of the text and determine appropriate anaphora.

## 4. CONCLUSIONS

In this paper the concept of anaphora resolution exercise is re invoked as it can play a very important role in the information extraction framework. The anaphora analysis is a process of finding hidden anaphoric contexts from the texts. The difficulty in extracting the anaphora prevails quietly in all linguistic processing frameworks, and it is too difficult for an unknown, non-domain texts. The concept of finding the anaphora in non-domain texts is proposed in this paper. Finding anaphora in typical contexts is described with an ad hoc process model. The uncertainty model and knowledge representation plays a glue role in the problem. The default logic is chosen for formalization of statements in order to find anaphora. Experiments are performed on Editorial pages of "The Hindu" daily.



**REFERENCES**

[1] Ariel Cohen, "Anaphora Resolution as Equality by Default", A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 44–58, 2007. © Springer-Verlag Berlin Heidelberg 2007.

[2] Ruslan Mitkov, "Anaphora Resolution", © Pearson Education, 2002, ISBN 0 582 32505 6.

[3] Pavel Makagonov, Konstantin Sboychakov, "Software for Creating Domain-Oriented Dictionaries and Document Clustering in Full-Text Databases", A. Gelbukh (Ed.): CICLing 2001, LNCS 2004, pp. 454-456, 2001. © Springer-Verlag Berlin Heidelberg 2001.

[4] Miroslaw Truszczyriski, "Modal Interpretations of Default Logic", Department of Computer Science, University of Kentucky, 1997.