International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 5– No.1, January 2013 – www.ijais.org

A Technique for Anaphora Resolution of Text

Vipin Kumar Pandey
Business Technical Analyst
Deloitte(U.S.)
Hyderabad

Shreya Solanki
Assistant Professor
AKG Engineering College
Ghaziabad

Kamini Sharma
Assistant Professor
ABES Engineering College
Ghaziabad

ABSTRACT
Information management is an important requirement in today's world. The anaphoric references hide the important information. The identification of anaphorically referred information is called as anaphora resolution which has significant impact on increasing the efficiency of information management, techniques including text summarization, information extraction etc. In this paper we have proposed a method for anaphora resolution to engineer information management. The proposed method acceptably determines the potential referents of the anaphora specially of the verb phrase form, distinguishes between pleonastic 'it' and the anaphoric 'it' and resolve the anaphora which is referred to after an interval of multiple sentences. The referents are stored in a list in the order of their occurrence in the discourse and eliminated from the list if they are not referred for to long. 'Recency' is used as a salience factor to select the correct referent if other information like gender, number and type are not suffices to estimate the correct referent for an anaphor. To achieve a more precise resolution system WordNet lexical database is exploited to compare the synonyms of the anaphor with its possible referents.

General Terms
Anaphora Resolution, Pronominal, Pleonastic Anaphor, Ontology based Anaphor.

Keywords
Anaphora, co-reference, antecedent, referent, WordNet.

1. INTRODUCTION
With the ever increasing growth of information resources, the demand for management of this information has increased. Anaphora or co reference resolution has its major application in summarization, information extraction, information abstractions, Natural language Understanding tasks and other such fields which play a vital role in information management activities. Thus anaphora resolution results to have a significant impact on knowledge management.

Language consists of collocated, related groups of sentences called discourse. The discourse model is the representation of the entities referred to in the discourse and their relationships. The significance of the concept of anaphora varies at different levels for different reasons. Its first significance is that it indicates the construction of discourse and its maintenance. Another, significance is the relationship it creates between various parts of the sentence which are syntactically related. Just another crucial reason for the study of anaphora is the problem of finding the reference to the anaphora in the discourse which is an intensive area of research in natural language processing. Anaphora also plays an important role in the area of linguistics related to cognitive science where it explains about the processing and understanding of language. Natural language processing involves building semantics out of unstructured text or spoken discourse. Resolving any referents in the discourse plays a crucial part in consolidating the meaning of text. For example in the sentence "Tom and Jack like mangoes and they eat them often", the meaning "Tom and Jack eat mangoes" can be extracted if the pronouns "they" and "them" is able to be resolved to the correct noun form.

Anaphora processing is a central topic in the study of natural language and has long been the object of research in a wide range of disciplines. Anaphora resolution is recognized as a very difficult problem in NLP [1]. For example in the sentence "Jessica likes playing tennis and she plays it often", the words "she" and "it" are anaphors that refer to "Jessica" and "tennis" respectively which can be easily understood by humans. However, developing an automated system to fully and correctly resolve the anaphora is not a straight forward approach, thereby undergoing intensive research. The elucidatory resolution of anaphora has also become increasingly important for several fields of real-world natural language processing, including machine translation, automatic text summarization, information extraction, and question answering. In this paper we have proposed a resolution system that works well for documents containing plain text. We exploit the features of such a genre of discourse which has relatively lesser number of subjective nouns.

2. ANAPHORA
2.1 Definition
The term Anaphora is used to define an expression that is used as a reference to another expression or entity in the discourse. For example in the sentence "John bought a car for himself", "himself" is an anaphora referring to its antecedent, "John". Anaphora resolution is the task of identifying referents or antecedents that might have been stated earlier or may be mentioned later in the discourse by an anaphor. Noun phrases, verb phrases and complete sentences can be the referents. Noun phrases can be definite or indefinite, a pronoun, a demonstrative or a reflexive [2].

Typically this problem can be divided into two parts:

(a) Finding the co-reference of a full NP (commonly referred to as co-reference resolution)

(b) Finding the reference of a pronoun or reflexive (commonly referred to as anaphora resolution).

2.2 Types of Anaphoric Expressions

The referents can be referred by different types of anaphoric references. These can be categorized into the following types:

2.2.1 Pronominal: This is the most common type where a referent is referred by a pronoun. For example in the sentence "Justin found the love of his life", the anaphoric expression 'his' refers to 'Justin'.

2.2.2 Definite noun phrase: The antecedent is referred by a phrase of the form "<the> <noun phrase>". Continued example, in the sentence "The relationship did not last long", where 'The relationship' refers to 'the love' in the preceding sentence.

2.2.3 Quantifier/Ordinal: The anaphor is a quantifier such as 'one' or an ordinal such as 'first'. Continued Example: "He started a new one" where 'one' refers to 'The relationship' (effectively meaning 'a relationship').

2.2.4 Pleonastic anaphor: is also known as null anaphor. In this category the pronouns "it," it's" and "itself" refer to nothing in particular. For example the "it" in the sentence "It might rain tonight".

2.2.5 Ontology based anaphor: This is the most difficult category of anaphor where the anaphor refers to some real world knowledge which has not been mentioned previously anywhere in the discourse. For example in the sentence "Jack bought an old car. The vehicle was in a good state", the word "vehicle" is such an anaphor which has no explicit relation as world knowledge such as that cars are vehicles is needed to resolve the reference [6].

2.2.6 Reader/Writer anaphor: the anaphor refers to the person consuming the discourse. In the case of an article it might refer to the reader. For example "you" in the sentence "If you pay peanuts you get monkeys".

3. RELATED WORK

Research in anaphora resolution has taken up pace with the advent of greater interest in knowledge engineering in recent years. In the realm of natural language processing, automated anaphora resolution has been an active area of research from past few decades. So far the research done under this area can be categorized into the following approaches:

a) Employ syntactic rules [4].
b) Allot certain salience weights to candidate referents [5] and [7].
c) Exploit statistical attributes of participating antecedents [3].

Any one or amalgamation of the above approaches is used by majority of the researchers to improvise their algorithms for a specific genre. It has been a complex problem to develop a universal algorithm which can work across many different or all the genres of Natural Language texts possessing specific traits and features. This is the reason that the research for resolution algorithms aim in the direction of specific genres. Other than the above mentioned approaches, the technology used for the implementation of these approaches like symbolic, neural networks, machine learning, etc. is another way of categorizing the research done under anaphora resolution.

Anaphora resolution approaches can also be broadly classified as knowledge-poor approaches and knowledge-rich, depending on the amount of contextual knowledge integrated into the system. Knowledge-rich approaches can be further divided into categories based on the type of knowledge employed [8]. First category is the syntax based approach and the earliest algorithm under this category was developed by Hobb[1977] making use of the fully parsed syntactic tree to find the antecedents, the results showed remarkable accuracy.

Second traditional knowledge-rich approach is discourse based approach using the Centering Theory (CT) to obtain the reference of pronoun which uses the salience of discourse entities and relates it to the referential continuity. This approach was used by Brenan, Friedman and Pollard [1987]. (BFP) using centering principle to rank prospective candidates. However the earlier approach of Hobb proved to be comparatively achieving higher accuracy than the approach employing CT for a particular genre of text. The CT-based approach presented with alteration called the Left-Right Centering approach (LRC). This psycholinguistic fact that listeners resolve references as soon as they hear was modeled in this approach. The LRC, if does not find the antecedent in first utterance, then antecedents in previous utterances are considered, going from left-to-right within an utterance.

Third, knowledge-rich approach was Corpus based used by Charniak, Hale. and Ge. in 1998 presented a statistical method for pronoun resolution based on the Hobb's algorithm. The Penn Wall Street Journal Tree-bank marked with co-reference resolution was used as a training corpus.

Fourth, knowledge-rich approach is the hybrid approach exploited by Lappin and Lease 1994 making use of more than one of the knowledge source including syntactic, discourse, morphological, semantic, etc. to rank potential antecedents in the discourse.

The other approach of anaphora resolution is knowledge-poor approach that makes use of the machine learning techniques is currently under the focus of researchers. The first such technique presented by Soon, Ng, and Lim in 2001 showed results comparable to the non- machine learning techniques, were able to resolve all definite descriptions.

The drive towards knowledge-poor and robust approaches was further motivated by the emergence of cheaper and more reliable corpus-based NLP tools such as POS taggers and shallow parsers, alongside the increasing availability of corpora and other NLP resources [3].

"A long-standing weakness in the area of anaphora resolution is the inability to fairly and consistently compare anaphora resolution algorithms due not only to the difference of evaluation data used, but also to the diversity of pre-processing tools employed by each system" [9]. Thus, the comparison among various anaphora resolution algorithms is

done on the basis of evaluation parameter MUC-6 and MUC-7. The current approaches are based on more sophisticated ML techniques like global models and kernel based approaches. Richer features like semantic information are exploited.

4. IMPLEMENTATION

4.1 Model Description

The implementation of the proposed system named as New Resolution System is described into the following steps:

4.1.1 Pre-processing: A given plain text document is parsed by the Stanford Lexicalized Parser and the result is output in a tree form. The complete input text has been converted into a single tree. The leaves of this output tree holds the part-of-speech tags which are determiner, conjunction, noun, proper noun, symbol, and all other part-of-speech in grammar. And the rest of the tree presents the relationship between words in every sentence and how these sentences are joined together to form the complete text. This result is given to the anaphora detector system.

4.1.2 Anaphora Detection and Resolution: The tree is scanned to extract the noun phrases, verb phrases and anaphora in the input text. Gender, number and type agreement is strictly followed to compare the anaphora with its possible referent and resolved when all these parameters match. Repositories of proper nouns such as male and female names are kept for gender agreement. Recency is taken into account to select the antecedent of an anaphor when other information like gender, number and type is found to be matching with more than one referent. These were the eliminative techniques to resolve the basic anaphors.

The next task performed by the algorithm is to find the action noun and verb phrase relationship wherever a verb phrases is an anaphoric expression. This is done by mapping the verb phrase and action-nouns. The verb phrases are converted into their basic form by using the WordNet. The synonyms of these basic verbs are then matched with the synonyms of the anaphora that potentially refers to it. If a match occurs this kind of anaphora is successfully resolved.

4.2 Algorithm

Step 1. First, parse the sentence file by the Stanford Lexicalized Parser which generates a tree with Part-Of-Speech tags held by the leaves.
Step 2. Find the verb-phrases, noun referents and Anaphora from the POS tagged tree produced in step 1.
Step 3. Collect all the referents in this sentence file in two parses of the tree description.
Step 4. Collect all the anaphora in the first of these parses.
Step 5. Assign integer value to the anaphora and (n+1)/2 to the referents.
Step 6. Gender, number and type are constraint variables that are strictly matched between the referent and anaphora.
Step 7. Gender agreement is implemented by using a list of proper nouns others than common words like ‘man’ and ‘woman’.
Step 8. Find the position in the text where the corresponding referent was last referred to.
Step 9. Remove the referent from the list that is not referred to for very long in the text (e.g. more than 20 sentences).
Step 10. Action-nouns are only compared against verb phrases and the proximity is used for detection if no other information is present.
Step 11. A list of action-noun mapped to the corresponding verb (e.g. reaction->react) is obtained using the list of proper nouns.
Step 12. Search the synonyms of these verbs from the WordNet lexical database.
Step 13. Verb head is first reduced to its basic form using the action-noun and verb pair list.
Step 14. Each verb phrase referent is checked whether its verb heads is one of these synonyms.
Step 15. If the anaphoric verb and referent verb are found to be synonymous then an arbitrary score is assigned.
Step 16. The anaphora referring to all the action, verb, nouns, and entities are obtained in the order of occurrence.

4.3 Design Constraints

The two necessary steps in the design of reference resolution are as follows:

a) Filter the set of possible referents by specific hard and fast constraints.
b) Set the preference for possible referents.
The constraints used for filtering the co reference are the following:

4.3.1 Number Agreement: Clear distinction between singular and plural referents. For example, in the sentence “Sally has a new dress. They are red.” The referent “they” is plural and does not refer to “dress” which is plural.

4.3.2 Gender Agreement: Male, female and non-personal genders are to be distinguished correctly by the resolution system

4.3.3 Person and case Agreement: The filter should precisely distinguish three forms of person. For example, in the sentence “You and I own Hondas. They love them.” the three forms are “You”, “I” and “They”. The system must interpret subject position, object position and genitive position.

4.3.4 Syntactic Constraints: The constraint that there should exist a syntactic relationship between a referring expression and its possible antecedent referent. For example in the sentence “Tom bought himself a chair”, “himself” refers to “Tom”. In another sentence “Tom bought him a chair”, “him” will not be resolved by the system to refer to “Tom”.

4.3.5 Selectional Constraints: For example in the sentence “Tom parked his car near the lake. He had driven it for hours.”, “it” refers to “car” not “lake” . This is the selectional constraint which is a restriction placed by the verb on its argument.

4.4 Technique Designing

The techniques that are used by the system to implement the above constraints precisely are:

4.4.1 Recency and Multi-Sentence Resolution

The entities introduced recently in the discourse are considered more salient than those introduced previously. For example in the sentence “Tom has Honda and John has Mercedes. Johanna likes to drive it.” “it” refers to “Mercedes” in accordance with ‘recency’. This is done by keeping a history list of the referents that are encountered in the in order

parsing of the discourse. The referents are appended as per the next sentence of the input text is parsed. To refrain from the inefficient and meaningless search for interpretation of indistinct anaphora, sporadically clipping of the list after every m words per n sentences is made. Those referents which are not referred another time by any anaphor or phrase for that defined span are discarded. This factor is implemented by keeping a variable to keep record of the latest reference to an antecedent. The resolved anaphora thus acts as noun phrases in the history list. This linking of anaphora to refer back to the preceding anaphora can be considered as chaining of anaphora [2].

4.4.2 Identify Pleonastic 'it'

The "pleonastic" 'it' works as a subject in a sentence which specifically has no meaning in itself but contributes in making the sentence grammatically correct. This type of 'it' is also called expletive or dummy pronoun and serves as empty subject markers [10]. For example, "It rains" is a sentence where "it" is a pleonastic pronoun which is not considered anaphoric since they do not have any antecedent but identifying such occurrences is important so that the coreference resolution system will not try to resolve them [11].

4.4.3 Resolving referents of verb phrase form

A verb phrase embodies an event or action which is referred to by the action-noun in a previously introduced sentence in the discourse. The action-noun refers to the noun that fall under the category of definite noun phrase coreference. This not so complicated example illustrates the use of verb phrases as referents: "Tom tried to encourage John to play the match. The effort was successful." 'The effort' here refers to the verb phrase in the preceding sentence, i.e. 'tried to encourage John to play'. This kind of verb phrases as illustrated in the above example can be resolved easily by finding out the pre-occurring definite noun phrase anaphora. Semantic or syntactic knowledge is required to resolve some other more complicated cases.

The design of the anaphora resolution system proposed here has been shown as an architectural model consisting of the sub-modules of the whole system in the Figure 1.

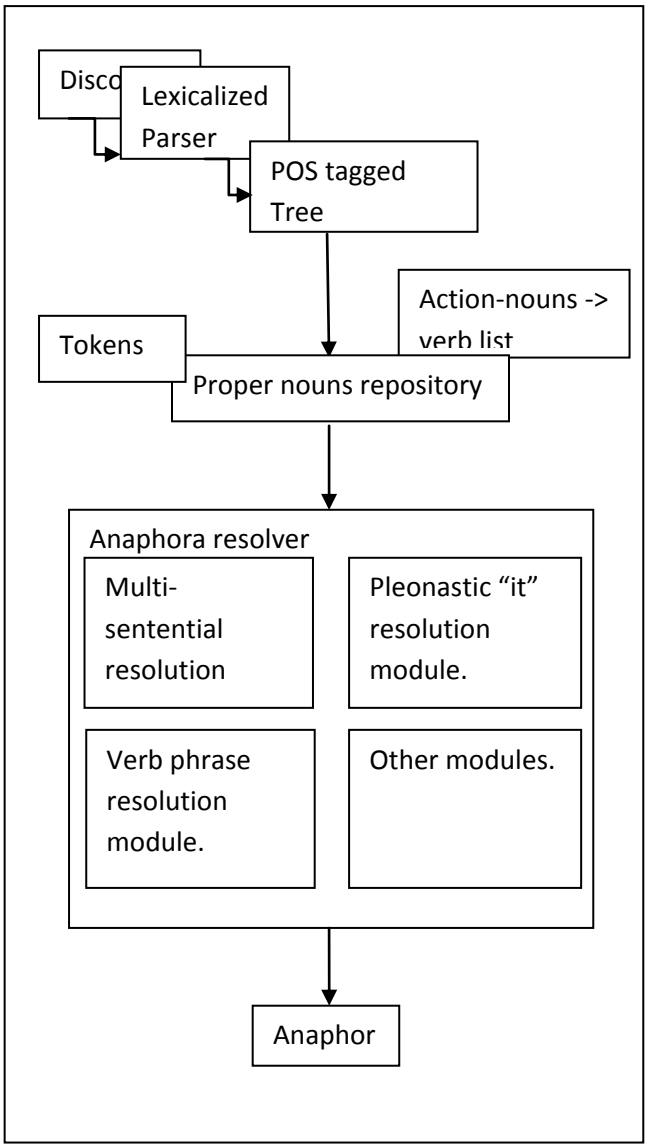

Fig 1: Modules of proposed anaphora resolution system.

5. Results

The idea of the system proposed by us in this paper is efficient to some extent however it has a limited scope. Sample text used for the measuring the correctness of the system consists of some general text. A standard corpus has not been used for this purpose because this resolver is genre specific and not comprehensive. The resolver has successfully resolved action-noun anaphora present in both the sample text. Similarly, the anaphora that is used for the antecedent introduced at a multi-sentential distance. Also, the resolver could extract the anaphora successfully whose reference is present in the same sentence by strictly following the gender, number and type agreement.

The following table summarizes the various characteristics of the two input sample texts.

The below table shows the number of resolved co references where the value in the brackets indicates the number of correctly resolved co references.

The precision, recall and F-measure of the New Resolver System using the above results of the summary are as follows:

Precision = 86.3%

Recall = 79.1%
F-measure = 82.3%

Table 1. Result of the New Anaphora Resolver System

Summary	**Sample Text 1** No. of Sentences = 25		**Sample Text 2** No. of Sentences = 31	
Sample text characteristics	**Actual Coref.**	**Resolved Coref.**	**Actual Coref.**	**Resolved Coref.**
No. of Proper Nouns	7	7	6	6
No. of Nouns	16	14[13]	20	20[17]
No. of anaphors	15	14[12]	23	20[17]
No. of pronominal anaphor	13	11[10]	9	8[7]
No. of definite noun phrase anaphors	2	2[1]	1	1[1]
No. of Verb Phrases	2	2[2]	1	1[1]
No. of pleonastic Anaphors	3	0	4	0
No. of action-noun Anaphors	1	1	2	2[2]

The table below summarizes the results of the comparison made among the existing anaphora/co reference resolver tools and the New Resolution System proposed by us in this paper.

Table 2. Comparison Results of Four Different Resolvers

Resolver Systems	**Percentage of Anaphora Resolved**	
	Sample Text 1	**Sample Text 1**
New Resolution System	90%	68.1%
GUITAR	69%	67%
JavaRAP	65%	54%
MARS	59%	51%

The results produced by our proposed system are noticeably better than the other resolution tools tested for the two sample texts as input data shown in the above table. The genre of the text is an important factor in determining the performance of any co-reference resolution tool. The same tool can give different accuracy result for different genre.

The user interface of the system has a load button that loads the text file from the disk and resolve button that functions to output the anaphora and its resolved referent present in the input text by first pre-processing with the Stanford Lexicalized Parser generating a POS tagged tree and then resolving.

A screenshot of the New Resolution System proposed in this paper shows the input text loaded by using the load button is given in Figure 2 and the corresponding and anaphora and resolving referent can be done by clicking the Resolve button of our API. We can save the output using the save option of our API into a file which can be used directly for any future computation.

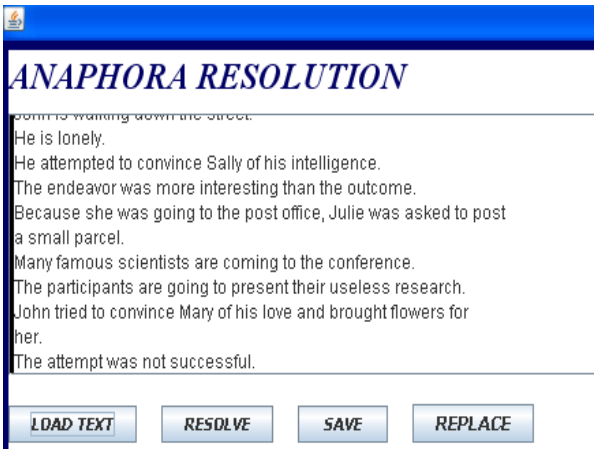

Fig 2. Input Text loaded

The anaphora and its resolved referent have been shown in Figure 3 by the screenshot of the output on the API. In the below Figure we are seeing that the texts we have loaded have been processed and the anaphora referent combination are shown with the help of -> symbol.

Initially in our approach we were seeing that some referents were coming wrong due to a long gap between anaphora and their referent due to which our program's efficiency was hampered.

So to improve on this factor we used Chaining technique which helps us to attain a greater accuracy rate. So for understanding the need of chaining we are elaborating it with an example. The text shown in the italics is the sample we have taken for explaining chaining.

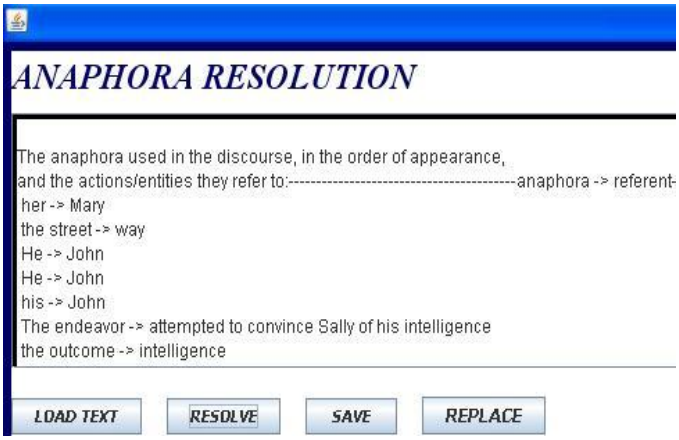

Fig 3:.Output with resolved referent

Mary saw a fat bald man in the park on her way.
John is walking down the street.
He is lonely.
He attempted to convince Sally of his intelligence.
The endeavor was more interesting than the outcome.
Because she was going to the post office, Julie was asked to post a small parcel.
Many famous scientists are coming to the conference.
The participants are going to present their useless research.
John tried to convince Mary of his love and brought flowers for her.
The attempt was not successful.

Here "the attempt" in the last sentence is referring to 4rth line. So if chaining will not be there then the results of resolution for "the attempt" will not be obtained. So in a Multi-sentential text like above anaphora resolution can be done through a process of 'chaining' and keeping a history list.

6. DISCUSSION AND FUTURE WORK

The anaphora resolution and other such methods provide the basis for extracting the important information from the source and thus providing a better platform on which the information management techniques can be exploited. There exists no best algorithm for fully automated anaphora resolution. However, different algorithms are proved to be accurate for different genres of text. The proposed algorithm deals successfully with the pronominal anaphora, definite noun phrase anaphora, verb phrase anaphora and pleonastic 'it'. The future direction of this work is to design a strategy to interpret the other forms of anaphoric expressions like associative action-noun anaphora. To find possibilities of such occurrences needs to be assessed and to find the techniques to resolve them. An evaluation mechanism to measure the performance of the system has to be designed. The performance measure of the technique of multi-sentential resolution with other such strategies can be determined. An optimal system for choosing the preference/score that is assigned to a more salient referent has to be established. Extended attribute knowledge can be gained by using adjective phrases or other such compositions. The attribute extraction feature if implemented will make it a more efficient in the interpretation of the anaphora based on world knowledge.

The NLP systems are desirable to be able to achieve the designing of such systems that are not language specific such as English, Spanish, etc. Another, interesting idea can be the designing of such systems which do not follow the customary chronological occurrence of events in the text because some discourse are such that they do not follow the chronological order. It becomes irrelevant to use this ordered approach for literature where old events are illustrated later in the discourse.

7. REFERENCES

[1] Denber, M. 1998.Automatic Resolution of Anaphora in English. Technical Report, Eastman Kodak Co. Imaging Science Division.

[2] Sayed, I.Q., Issues in Anaphora Resolution. Unpublished.

[3] Barbu, C. and Mitkov, R. 2001 Evaluation tool for rule-based anaphora resolution methods. Proceedings ACL, 39th Annual Meeting and 10th Conference of the European Chapter, France, 2001, pp 34-41.

[4] Baldwn, B. 1997. CogNIAC: High precision coreference with limited knowledge and linguistic resources. Proceedings ACL '97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution, Madrid, Spain, 1997, pp. 38-45.

[5] Ho, H., Min, K., and Yeap, W. 2004. Pronominal Anaphora Resolution using a shallow meaning representation of sentences. Proceedings of the 8th Pacific Rim International Conference on AI, 2004, pp 862-871.

[6] http://nuigalway.academia.edu/SiegfriedHandschuh/Papers/568236/Ontology-based_linguistic_annotation

[7] Mitkov, Y. R. 1998. Robust pronoun resolution with limited knowledge. Proceedings COLING'98/ACL'98, Montreal, Canada, 1998, pp. 869-875.

[8] Deoskar, Tejaswini. Techniques for Anaphora Resolution: A Survey. Unpublished.

[9] Mitkov, R. Outstanding issues in anaphora resolution.

[10] Haegeman, L. 1991. Introduction to Govemment and Binding Theory, Blackwell Publishing, 1991, pp 62.

[11] Mitkov, R., Daarc., Branco, A., McEnery, T., and Lisbon,. 1955. Anaphora Processing: Linguistic, Cognitive And Computational Objectionable web content classification using Neural Network. vol. A247, Phil. Trans. Roy. Soc. London, April 1955, pp. 529–551.