International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 6, June 2016

A Survey of Anaphora Resolution Approaches for Different Indian Languages

[1] Harjinder Kaur, [2] Kamaldeep Kaur
[1]PG Student,[2]Assistant Professor
Department of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India,
[1] kaurharjinder28@gmail.com[2]kapilsharma701@gmail.com

Abstract: **— Anaphora Resolution is a key problem in the field of natural language processing. It is the problem of identifying the referents in the discourse. Pronominal anaphora resolution is a subpart of anaphora resolution where pronouns are referred to noun antecedents. This paper describes the research done in the development of anaphora resolution systems for various Indian Languages.The process of anaphora resolution can become increasingly complex when we encounter complex sentences. The aim of study is to resolve the ambiguity from complex sentences using semantic and pragmatic knowledge.The problem tackled by many Indian languages is no capitalization, lack of standardization, requirement of Unicode based tools. It can be concluded that various approaches are used to resolve anaphors for different Indian Languages.**

***Index Terms*—Natural Language Processing (NLP), Anaphora Resolution (AR), Part of Speech (POS) Tagging, Noun Phrase (NP).**

I. INTRODUCTION

Anaphora Resolution is a process of referring back to the previous element in the discourse. Discourse is a group of inter-related sentences. The pointing back word is known as anaphor. The entity to which anaphor refers is its antecedent. The process of mapping the referring expression to the co correct antecedent in the discourse is called anaphora resolution.

Among the various types of anaphora such as pronominal anaphora, noun anaphora, noun phrase anaphora and zero anaphora, pronominal is the most common type of anaphora.

All NLP applications such as machine translation, automatic summarization, question answering system etc., require successful identification and resolution of anaphora. Resolution of anaphora is based on various factors such as regency, animistic, person and case, number and gender agreement. Anaphora resolution is a challenging task in NLP because it need not only knowledge but also expertise of all language processing domains such as morphological and lexical knowledge, syntactic knowledge, semantic knowledge, discourse knowledge, real world knowledge. Sometimes pragmatic knowledge is very important in anaphora resolution [4], [7].

II. TYPES OF ANAPHORS

The different types of anaphors according to the form of the Anaphor are as follows [1]: Nominal Anaphora: Nominal anaphora is used when a referring expression (pronoun, definite noun phrase or proper name), has a non-pronominal noun phrase as its referent.

- _ Pronominal Anaphora: Pronominal anaphora is the most common type of anaphora. Pronominal anaphora resolution is a process where pronouns are referred to the noun referents.
- _ Lexical noun phrase Anaphora: Lexical noun phrase anaphora is realised syntactically as definite noun phrase, also called definite descriptions, and proper names.
- _ Zero Anaphora: Zero anaphors are known as invisible anaphors. It is described as referring back to an expression that provides the information necessary for interpreting the gap.
- _ One Anaphora: One-anaphora is the case when the anaphoric expression is described by a ”one” noun phrase.
- _ Intrasentential Anaphora: In this case, anaphora and its antecedent are present in the same sentence.
- _ Intersentential Anaphora: In this case, antecedent is in a different sentence from anaphora.

III. APPROACHES TO ANAPHORA RESOLUTION

The various approaches to resolve anaphora are as follows:

Rule based approach : Small set of rules are created to identify the antecedents of NPs of interest. System does not required training [1]. Lakhmani et al.[3] proposed anaphora resolution system for Hindi language. In this, authors performed two experiments on two different types of data sets to resolve the anaphors. The first experiment used text from a childrens story. Another experiment used text from news articles. The experiments based on anaphora resolution have been conducted using some constraint sources which will form the base of anaphora resolution task. On the basis of experimentation, results are obtained which shows final accuracy of approx 71%. Liang et al. [9] proposed Automatic Pronominal Anphora Resolution in English Texts. In this, the anaphora resolution has been achieved by using Word Net ontology and heuristic rules. The system identifies both inter-sentential and intrasentential antecedents of anaphors. The evaluation task is based on random texts selected from the Brown corpus of different genres. There are 14,124 words, 2,970 noun phrases and 530 anaphors in the testing data. The overall success rate reaches 77%.

Mohan et al.[12] proposed a Rule Based Anaphora System for Telugu Language. In this, some syntactic cues for each Telugu pronoun such as Personal, Demonstrative, Indefinite, Interrogative , Reflexive etc has been made and based on these syntactic cues authors presented some rules for the pronominal resolution. The system was evaluated on limited set of data. The results depend mainly on the gender agreement. It is generating 58%, 57%, 80% and 48% accuracy for personal, Demonstrative, Interrogative and reflexive pronouns respectively. The total accuracy of the system is 60.75%. Dutta et al.[16] proposed a Pronominal anaphora resolution system for Hindi Language using Hobbs Algorithm. In this, a modified version of Hobbs Naive Algorithm for Hindi has been implemented using the free word order and grammatical role in pronoun resolution in Hindi. This algorithm is used as baseline algorithm which provides syntactic information instead of semantic information. The algorithm is used for the roles of subject and object and their impact on anaphora resolution for reflexive and possessive pronouns. The algorithm has been implemented for limited set of sentences defined by grammar.

This algorithm uses grammatical features, which can be used to resolve pronouns. Pralayankar et al.[20] proposed a Sanskrit Analysis System (SAS) which is used to resolve pronominal anaphors in Sanskrit. In this, an algorithm is developed to resolve the anaphors from the input text. Using this algorithm, POS tagging has been used to assign the tags to the different words. In this, possible anaphora-antecedent pairs are made.

The sentences which can't be analyzed are excluded. According to the algorithm, the anaphors has been categorized. Several modules of SAS have been developed and has been used in testing and evaluation phase. _ Corpus based approach: Uses selection patterns, statistics or co-occurrence patterns observed in the corpus. Training (Machine learning) is required [1].

Pal et al.[4] introduced the issues related to syntactic / semantic structure of Hindi and influence of cases on pronouns, mainly personal pronoun. In this, EHMT (English-Hindi

Machine translation is demonstrated to substantiate the need of anaphora resolution for NLP applications. The authors described the comparison of several translation systems for Spanish and English languages using different data sets. This translation system gives good reported results for Spanish and English languages. The reported success rate for Spanish and English language is 80.4% and 84.8 % respectively. Kamune et al. [5] proposed an algorithm which uses the combination constraint-based and preferences-based architectures for resolving anaphors. It is used to identify intersentential and intra-sentential antecedents of Third person pronoun anaphors , Pleonastic it and Lexical noun phrase anaphora. This algorithm at first defines an anaphoric space, then applying constraints and finally applying preferences. The output is generated by Charniak parser (parser05Aug16) and salience measures derived from parse tree. Devi et al.[8] used a generic anaphora engine for Indian Languages .The authors analysed the similarities and variations between pronoun and their agreement with antecedents in

Indian Languages. The algorithm developed uses the morphological richness of Indian languages. The machine learning approach uses the features which can handle major Indian languages. The system has been tested with Indo-Aryan and Dravidian languages such as Bengali ,Hindi and Tamil. Deepa K et al. [13] proposed a System for Tamil Anaphora Resolution(STAR). In this, a decision tree based machine learning algorithm is used for anaphora resolution. The system first identifies which are pronouns and noun phrase for the received input text. The system addresses the addition challenges in Tamil like morphological richness, semantic ambiguity and noun phrase chunking. The system built a feature vector lexical, syntactic and semantic features for machine learning. The experiments has been done on NER corpus(Tourism) of TDIL dataset. The text contains 10,000 sentences and totally there are 2295 pronouns out of which the selected portion of training data contains 239 pronouns. The system has obtained f-measure of 73% and up to 77% accuracy.

Ram R et al. [15] presented pronominal resolution in Tamil using Tree CRFs, a machine learning technique. Tamil is a morphologically rich language. The features for learning has been developed by using the morphological features of the language and from the dependency parsed output. The learning features has been used in the salience factor approach and the constrains mentioned in the structural analysis of anaphora resolution. The work has been carried out on tourism domain data from the web. It is generating 70.8% precision and 66.5% recall.

Singh et al.[17]presented Sensitivity Analysis of Feature Set used for Anaphora Resolution. Sensitivity Analysis is the process which is used for doing systematic review for a sequence of parameter, feature set and decisions are used to calculate the impact of these parameters on the study. In this, authors consider two features out of seven tags which were used to resolve the anaphora in Hindi. These tags and their values had been analyzed for the corpus. Authors analyzed 165 news items of Ranchi Express from EMILEE corpus of plain text. It consists 1745 sentences. Sikdar et al.[18]introduced an anaphora resolution system for Bengali language. In this, number of models have been developed based on the heuristics used and the particular machine learning employed. A series of experiments are performed
for adapting domain(BART) for Bengali. The work is divided into two folds: (i) an attempt to build a machine learning based anaphora resolution system for a Indian language and (ii)domain adaptation of a state-of-the-art English co reference resolution system for Bengali, which has completely different orthography and characteristics. The system produces the recall, precision and F-measure values of 56.00%, 46.50% and 50.80%, respectively.

Mehla et al.[21] introduced a machine learning approach to resolve event anaphora in Hindi language. Event anaphora are those pronouns which refers to events, the possible candidate referents are verbs, clauses and prepositions. In this, an algorithm is proposed for resolving event anaphora and accuracy of algorithm is checked for different algorithms. The algorithm works very well for large number of sentences. The testing data contains 1071 concrete pronouns. The remaining data two-third data is used in 4-fold iteration for training and development. The accuracy of the algorithm on different algorithms is quite impressive.

Knowledge poor approach: This approach uses domain and linguistic knowledge (which is difficult to represent and process) require considerably human inputs. This approach does not require linguistic knowledge therefore, is less labor intensive and less time consuming. This approach is difficult to implement when there is lack of annotated corpora[1]. Dakwale et al.[14] introduced a hybrid approach for anaphora resolution in Hindi. In this approach, dependency structures has been used by a rule-based module to resolve simple anaphoric references, while a decision tree classifier helps to resolve more ambiguous instances, using grammatical and semantic features. The Treebank data set is used for training and testing. The training data contains 2162 entity pronouns and test data contains 1071 entity pronouns The accuracy of rule based system for different types of pronoun is 60%. The overall performance of the hybrid system achieved over the rule based system by using different set of features is 70%. Sobha et al.[19] introduced a approach using Named Entity and ontology. In this, semantic disambiguation of the antecedent and anaphor has been attained by using a Semantic Disambiguator. The Semantic Disambiguator helps to resolve the issues related to animacy and real world identity of the nominals (NP) and thus helps to proposed the most likely candidate antecedent for an anaphor. The base system uses salience factors and salience weight of the candidate NPs for identifying the antecedent from the list of possible candidates for antecedent-hood. The approach have been tested on ACE2008 documents.In this, 200 documents have taken from the UseNet domain, which contained 1077 pronouns. It is generating 80% precision and 74% recall.
_ Discourse based approach: Discourse based resolution approach is an attempt to exploit discourse structure, specially the relationship between references and discourse theme, to resolve definite references[1].

Singh et al.[7] proposed Anaphora Resolution System for English language. In this,two models has been proposed for anaphora resolution. Resolution of anaphora is based on two factors which is Recency factor and Animistic Knowledge. Recency factor is implemented by using Lappin Leass approach in first model and using Discourse based approach in second model. Information about animacy is obtained by Gazetter method. The result given by these two models gives 70 to 80 percentage accuracy respectively.

IV. METHODS FOR ANAPHORA RESOLUTION

The various methods to resolve anaphora is as follows [2]: _ Centering Theory: Centering theory is based on discourse based approach. This theory is used to model what a sentence is speaking about a framework. The framework can be used to identify that pronouns are referring to which entities in a given sentence. Attentional salience of discourse entities are modelled in this theory.

From the computational point of view centering based approaches are more attractive. It is because they obtained the required information from the properties of utterances alone. Lappin Leass algo: Lappin Leass algo is a hybrid method which is used to assign the assign tha

salience values to a sentence on the basis of their salience factors.

Different weights are given to the factors on the basis of their relevancy. Gazetter method: Gazetter method is used to provide the external knowledge to the system by creating lists.

On the basis of certain operations elements of lists are created. It is also known as List Lookup Method. Lappin et al.[6]proposed an algorithm which is used to identify the noun phrase antecedents of third person pronouns and lexical anaphors. Authors performed a blind test of this algorithm on computer manual texts and manual text containing 360 pronoun occurrences. This algorithm successfully identifies the antecedent of the pronoun for 86% of the pronoun occurrences. In this, semantic and real-world knowledge apply to the output of an algorithm that resolves pronominal anaphora on the basis of syntactic measures of salience, regency and frequency of mention.

Lakhmani et al. [10] used Gazetteer Method for Resolving Pronominal Anaphora in Hindi Language. The resolving system used Recency as a baseline factor for resolving anaphora. Animistic knowledge is used to differentiate between animate and inanimate things. In this, experiments are performed on different data sets having different style of written text in Hindi language each having 100 to 300 words. The system gives approximate 60 to 70 percentage successful identification of anaphora.

Singh et al. [11] presented a Comparative Analysis of two resolution systems. In this, two computational models uses Gazetter method for resolving anaphora in Hindi language. Different classes of elements are created in Gazetter method. The two models use Recency and Animistic factor for resolving anaphors. The experiments were conducted on Hindi stories, news articles and biography content from wikipedia.

The data set contains 100 to 300 words. The accuracy of computational model depends on the type of input data. The approximate accuracy of both the model is compared. The system gives approximate 70 % accuracy.

V. DISCUSSION

Anaphora resolution Approaches are divided into four categories. First is rule based approach which was implemented by Lakhmani et al.[3] , Liang et al. [9] , Mohan et al.[12], Dutta et.al [16], Pralayankar et al.[20] requires number of natural language resources e.g. part of speech tagger and parser. Corpus based approach is the second one which exhibits the features of the genre to resolve anaphora resolution and it is implemented by Pal et al.[4], Kamune et al. [5], Devi et al.[8], Deepa K et al. [13], Ram R et al. [15], Singh et al.[17], Sikdar et al.[18], Mehla et al.[21].Knowledge poor approach which does not rely on linguistic and domain knowledge and it is used by Dakwale et al.[14], Sobha et al.[19].Fourth is discourse based approach which is modeled through a sequence of utterances and is implemented by Singh et al.[7].

Table I
Status Of Research Done In Indian Languages

Languages	Authors
Hindi	PriyaLakhmani et al.2013,Dutta et al. 2012, Mathur et al.2014, Dakwale et al.2013 and Kamlesh and Prakash2008
English	Kamune and Agrawal2015, SmitaSingh et al.2014,Liang and Wu2003
Tamil	Deepa and Deisy2016, Ram and Devi2013
Bengali	Sikdar et al.2013
Sanskrit	Pralayankar et al.2008
Telugu	Mohan and Sadanan-dam2014

VI. CONCLUSION

In this study, we have reviewed various approaches, which are used to resolve the anaphors correctly for different Indian Languages. It can be concluded that most of the approches of anaphora resolution are inspired from other languages such as English and European languages and hence, are adopted for Indian languages. The problem tackled by many Indian languages is no capitalization, lack of standardization, requirement of Unicode based tools The task of anaphora resolution is done for various indian languages.The anaphora resolution task seems quite difficult for complex sentences.In this paper, four approaches are used to resolve the ambiguity from complex sentences using semantic and pragmatic knowledge.

REFERENCES

[1] D. S. Yadav, K. Dutta, P. Singh, and P. Chandel, "Anaphora resolution for indian languages: The state of the art."

[2] A. J. Komal Mehla and Karambir, "Event anaphora resolution in natural language processing for hindi text," International Journal of Innovative Science, Engineering and Technology, 2015.

[3] P. Lakhmani and S. Singh, "Anaphora resolution in hindi language,"
International Journal of Information and Computation Technology. 2013.

[4] T. L. Pal, K. Dutta, and P. Singh, "Anaphora resolution in hindi: Issues and challenges," International Journal of Computer Applications, vol. 42, no. 18, 2012.

[5] K. P. Kamune and A. Agrawal, "Hybrid approach to pronominal anaphora resolution in english newspaper text," International Journal of Intelligent Systems and Applications (IJISA), vol. 7, no. 2, p. 56, 2015.

[6] S. Lappin and H. J. Leass, "An algorithm for pronominal anaphora resolution," Computational linguistics, vol. 20, no. 4, pp. 535–561, 1994.

[7] D. P. M. Smita Singh, Priya Lakhmani and D. S. Morwal, "Analysis of anaphora resolution system for english language," International Journal on Information Theory, 2014.

[8] S. L. Devi, R. V. S. Ram, and P. R. Rao, "A generic anaphora resolution engine for indian languages." in COLING, 2014, pp. 1824–1833.

[9] T. Liang and D.-S. Wu, "Automatic pronominal anaphora resolution in english texts." in ROCLING, 2003.

[10] S. S. Priya Lakhmani and D. P. Mathur, "Gazetteer method for resolving pronominal anaphora in hindi language," International Journal of Advances in Computer Science and Technology, 2014.

[11] D. P. M. Smita Singh, Priya Lakhmani and D. S. Morwal, "Comparative performance analysis of two anaphora resolution systems," 2014.

[12] D. C. Mohan and M. Sadanandam, "Telugu pronominal anaphora resolution."

[13] K. Arul Deepa and C. Deisy, "Anaphora resolution in tamil through decision tree learning-challenges and tackles," Int J Adv Engg Tech/Vol. VII/Issue I/Jan.-March, vol. 533, p. 538, 2016.

[14] P. Dakwale, V. Mujadia, and D. M. Sharma, "A hybrid approach for anaphora resolution in hindi." in IJCNLP, 2013, pp. 977–981.

[15] R. Ram and S. L. Devi, "Pronominal resolution in tamil using tree crfs," in Asian Language Processing (IALP), 2013 International Conference on. IEEE, 2013, pp. 197–200.

[16] K. Dutta, N. Prakash, and S. Kaushik, "Resolving pronominal anaphora
in hindi using hobbs algorithm," Web Journal of Formal Computation and Cognitive Linguistics, vol. 1, no. 10, pp. 5607–5607, 2008.

[17] P. Singh and K. Dutta, "Sensitivity analysis of feature set employed for anaphora resolution," International Journal of Computer Applications, vol. 128, no. 14, pp. 10–14, 2015.

[18] U. Kumar Sikdar, A. Ekbal, S. Saha, O. Uryupina, and M. Poesio, "Anaphora resolution for bengali: An experiment with domain adaptation," 2013.

[19] L. Sobha, "Anaphora resolution using named entity and ontology," in Proceedings of the Second Workshop on Anaphora Resolution (WAR II), Ed Christer Johansson, NEALT Proceedings Series, vol. 2, 2008, pp. 91–96.

[20] G. N. Jha, L. Sobha, D. Mishra, S. Singh, and P. Pralayankar, "Anaphors in sanskrit," in Proceedings of the Second Workshop on Anaphora Resolution, vol. 2, 2008.

[21] A. J. Komal Mehla and Karambir, "A machine learning approach to resolve event anaphora," International Journal of Scientific Engineering and Applied Science.