International Journal of Management, IT & Engineering
Vol. 7 Issue 12, December 2017,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A RELATIVE STUDY OF FACTORS AND APPROACHES FOR HINDI ANAPHORA RESOLUTION

Seema Mahato [*]
Ani Thomas [**]
Neelam Sahu [***]

Abstract

A literature survey on theoretical approaches of anaphora resolution system and various factors among which these approaches are based is presented in this paper. From last two decades, the significant amount of work on Natural language processing application such as anaphora resolution has been gradually increased but with respect to Indian language, especially Hindi, there is a long way to go. The success rate of anaphora resolution system also improves the performance of machine translation system, automatic text summarization and question answering system, etc. Many researchers implemented different factors suitable to their approach for anaphora resolution among which are gender agreement, number agreement, semantic analysis, salience measurement, word sense disambiguation, etc. The paper reviewed the factors useful for resolving anaphora and incorporation these factors for building the strategies for resolution and computational behavioral of anaphora resolution approaches. The authors finally put forward the lessons learnt from it, which may prove to be important knowledge sources for different information retrieval tasks.

Keywords:

Anaphora;
Anaphora resolution approaches;
Computation strategies;
Natural language processing;
Hindi language.

[*] **Research Scholar, Dr. C.V. Raman University, Bilaspur, Chattisgarh, India**

[**] **Head & Professor, Dept. of IT, Bhilai Institute of Technology, Durg, Chattisgarh, India**

[***] **Professor, Dept. of IT, Dr. C.V. Raman University, Bilaspur, Chattisgarh, India**

1. Introduction

An anaphora is a proform that refers to an entity comes into existence prior to it and its resolution deals with the identification such entities which is a difficult task. The entity to whom an anaphor refers is its antecedent. The anaphor may have clause, verb phrase, sentences or part of discourse as antecedent (Mitkov, 2002). Depending upon the comprehension of a discourse or corpus, the identification process has been done at clause level, utterance level, or discourse level (Delmonte et al., 2006). Anaphors which have antecedents are called as anaphoric and those does not have are non-anaphoric. Anaphora can be categorized as definite noun phase anaphora, one-anaphora, nominal anaphora, pleonastic 'it' anaphora, zero anaphora, pronominal anaphora, and lexical anaphora (Yadav et al., 2016). Before the resolution the discourse or corpus has to annotate manually or automatically which provide the morphological information. Based on some approach the anaphora resolution (henceforth AR) begins, it analyzes each sentence individually for identifying the referent or antecedent of an anaphora and expands its scope beyond this sentence if the antecedent is not found. The major tasks in AR process are identification of anaphoric and non-anaphoric pronouns in discourse too. Not all the researchers perform this task which may degrade the overall result. Though there are similarities in these two identification, but there are significant differences in the function of pronouns whether it precedes the referent or not. All noun phrases preceding an anaphor, whether intra-sentential or inter-sentential considered as potential candidates and placed into a set of possible candidates for being an antecedent.

Potential candidate have been chosen on the basis of number features selection attributes. Different constraints and preferences are used to formulate the pair of anaphor and its correct antecedent. Some of the strong constraints and preferences are discussed in this paper. The noun phrases which does not get fit into the criteria of defined constraints and preferences or factors are removed from the set. The scope for searching potential candidates for antecedents may vary from 3-17 sentences back or to the left of anaphora, which depends on the implementation of algorithm and the approach on which they are based. Finally the most appropriate noun phrase is picked from candidate set by employing statistical approach. The various approaches discussed in the paper tried to resolve various issues of anaphora resolution but none of the existing theory or methodology claims to resolves all types of anaphora. Researchers have used rule or corpus

based approaches and few of them combined these with machine learning techniques to improve the performance. These approaches are may be cost-effective, fast or highly performable depending upon the extent of semantic and/or syntax knowledge utilization and additional domain knowledge. Researchers have used globally recognized different metrics apart from F-measure, which are behaving distinctly such as MUC, B3, BLANC, CEAF, etc. for measuring the performance of anaphora resolution (Moosavi et al., 2016).

2. Factors for anaphora resolution

Mitkov (1997) has suggested factors for anaphora resolution useful for any natural languages. The most common factors that are taken into account during anaphora resolution for filtering out the unwanted candidates and to generate the list of potential candidates of an antecedent are listed here. The authors tried to discuss these factors with the support of examples in Hindi text so that it could help researchers for building efficient resolution algorithms for Hindi.

2.1 Gender and number agreement

Morphological knowledge provides information regarding gender (feminine, masculine, and neuter) and number (singular, plural). It is seen that gender and number agreement play important role in resolution. These heuristic factors help in choosing noun phrase as potential candidate for being an antecedent. Generally AR checks for resemblance in gender and number attributes of both anaphor and the potential candidate (Lakhmani et al., 2013). If they don't match, the candidate is filtered out otherwise taken into consideration and is placed in the list of potential candidate. Like noun phrase having the same gender and number is considered as potential candidate. Researchers not compulsorily employ these factors. It is easier to identify the gender of a pronoun in English language but in case of Hindi language, the gender of a pronoun is determined by its nearest verb phrase.

(1).“गीता और मनोज बाजार गये. वह बाजार से बहुत सारा खिलोना लाई”.

For example in the sentence (1), वह pronoun refers to गीता as the verb phrase indicates feminine comprehension. After a little change in verb phrase in the above sentence: “गीता और मनोज

बाजार गये. वह बाजार से बहुत सारा खिलोना लाया", the same वह refers to मनोज. In Hindi वह refers to singular entity whereas वे refers to plural entities which can be seen in the sentence "गीता और मनोज बाजार गये. वे बाजार से सारा खिलोना लाऐ" where the referent of वे are both गीता and मनोज.

2.2 Semantic analysis

Sometimes morphological, lexical and syntax knowledge are not enough to factorize the candidate. Moreover, semantic analysis can be used to proceed further. Semantic analysis checks the animacy and named entity category of anaphor and the antecedent (Sharma et al., 2017). The anaphor and the antecedent which do not have semantic feature match are filtered out. WordNet and dictionary entries are rich sources for providing lexical and semantic knowledge (Jain et al., 2013). These resources are generally hand-coded. In other words, it provides metadata about a word that a human brain commonly sense from comprehension. Lexical knowledge provides synonymy, hypernymy and meronymy of words. Lexical relations are very much dependent on the context/comprehension. Synonym set of each word is mapped to its ontology (hierarchical construct). The leaf node ontology is the word and top node contain syntactic category of a word. Further, the relations between different synonym sets are seized. Ontology of few Hindi words like बाज, बरगद, बहू, and कार generated by *Hindi WordNet*[1] (Jain et al., 2013) are graphically depicted below in figure 1.

[1]http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

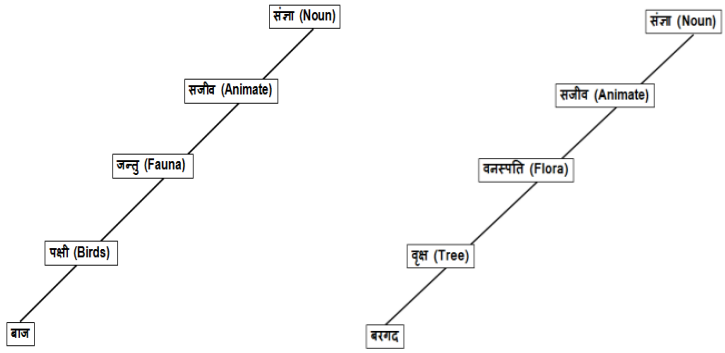
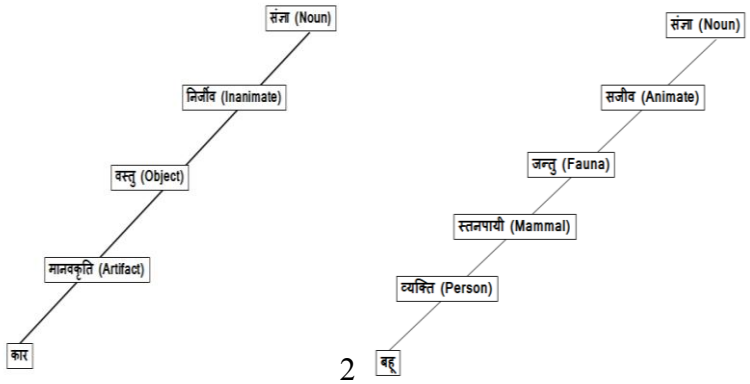


Figure 1(a) Ontology of words **'बाज and बरगद'**



2

Figure 2(b) Ontology of words **'कार and बहू'**

The most explanatory node of each ontology is the leaf node which contain the word ('बाज, बरगद, कार, बहू') and moving up in the hierarchy, each node syntactically turn up into more generalized form which can be represent as सजीव (Animate) -> जन्तु (Fauna) -> पक्षी (Birds) for बाज and so on.

Consider the sentences (2a) and (2b). In both the sentences, the subject is "संजय" and the object is "अनिल". The pronoun "उसने" in the embedded clause of both the sentence does not compulsorily refer to the subject of the sentence. From examples (2a) & (2b), prerequisite of verb semantics in anaphora resolution could be realized. In (2a), उसने is referring to संजय (a subject) whereas in (2b) it is referring to अनिल.

(2a).संजय ने अनिल को बताया की उसने बिल घुमा दिया.

(2b).संजय ने अनिल को डांटा क्योकि उसने बिल घुमा दिया.

2.3 Salience measurement

Selection of antecedent from set of potential candidate relies on too many parameters like which candidate is recently used or introduced prior to anaphor, whether a candidate is head noun or not, how frequently a candidate is repeated, if a candidate is part of sentence whose construction

is similar to other sentence, emphasis of verb, etc. These parameters are named as recency, subject emphasis, mention frequency, etc. (Lappin et al., 1994). These parameters are employed after implementation of syntactic agreements on set of potential candidate. Each parameter is assigned with a value and integrated sophistically in an algorithm crafted for ranking the candidates on the basis of their emphasis or saliency. Depending on the consolidated highest value or rank, a candidate is chosen as actual antecedent of an anaphor. The presence of salience measures for anaphora resolution necessarily improves the performance. In the example "मुकेश ने प्लेट गिरा दी. वह जोर से टूट गया", वह is a demonstrative pronoun and refers to प्लेट instead of मुकेश because प्लेट is more salient as introduced recently than मुकेश.

2.4 Word sense disambiguation

Word sense disambiguation (henceforth WSD) determines the sense or meaning of a word in a sentence when it has multiple meanings in ontology (Surekha et al., 2016). The meaning that has most sense with regard to context has to be select. Consider an example below of a Hindi word **"सोना"** with all its possible senses and their use in Hindi sentences (3a) and (3b). In (3a) and (3b), सोना fit into the sense2 and sense1 respectively.

Example: Possible senses of सोना

Sense 1: सोना [gold], धातु [metal],रंग[color], सिक्का [coin]

Sense 2: सोना [sleep], नीद [sleep], सो[to sleep], रात [night], दौरान [during]

(3a).मैं बहुत थक गया हू . मुझे अब सोना चाहिए.

(3b).जायदाद मे रीना को पैसे के अलावा सोना चाहिए.

WSD system more efficiently pick optimal senses for every noun and even for repeated noun in the sentence from trained data with respect to discourse which improve the anaphora resolution through accurate information retrieval (Navigli, 2009).

2.5 Named entity recognition (NER)

NER system is used in preprocessing phase as one of modules in anaphora resolution to provide semantic knowledge about named entities or proper nouns and to recognize and categorize them in a particular class such as person, place, date, organization, river, sport, location or geo-political entities occurring in a discourse (Nadeau et al., 2006). It does not make any difference between feminine and masculine entities. It helps in anaphora resolution by determining that which two phrases belongs to the same class.

For example, अमित/*Person* साईकल/Artifact से स्कूल/*Organization* जाता है.

But there are few issues related to NER such as one proper noun may occur in multiple forms like UK is उत्तराखंड or United Kingdom, जर्सी word is talking about the state "New Jersey" or a kind of cow, May is a month or verb, "IT" is a city in Mississippi State or a pronoun, etc.

3. APPROACHES IN ANAPHORA RESOLUTION

Mitkov (1999) on basis of computational strategies categorized AR approaches as traditional, alternative and knowledge poor. Most of the approaches under "traditional" or "alternative" use heavy linguistic and domain knowledge, developing which means that a remarkable amount of resources (like time, computational power, manual input of people) have to be taken into consideration which in turn is very labor-intensive. So the interests of researchers towards corpus-based and knowledge-poor approaches have been seen which are quite inexpensive and impervious to failure and more over offers only approximate solutions. So Mitkov (1999) have suggested knowledge-poor approach as one more approach apart from knowledge rich. Figure 2 graphically represents all the different approaches for resolving anaphora. Figure 2 graphically represents all the different approaches for resolving anaphora.

3.1 Traditional or knowledge-rich approach

The knowledge-rich approaches intakes the manually preprocessed data which generally includes manual removal of pleonastic pronoun 'it' which is quite a labor-intensive task. The immense amount of syntactic and semantic information of each word in the discourse acts as a resource for knowledge database. The parser employs these data for analyzing and resolving the anaphors. These approaches are mainly based on some existing theories and rely on of immense syntax and semantics knowledge sources but avoided domain knowledge like Lappin and Leass' syntax-

based approach (Mitkov, 1999). The factors for AR discussed above are used in this technique for making set of potential references. Domain, discourse or heuristics knowledge are heavily accompanied in such approach. All candidates are considered as equal but the final decisions for selecting the best of the remaining is based on the combined score attached to each and how a candidate is credible. Evaluation of these approaches was typically carried out by hand on a small set.

3.2 Alternative approach

Alternative approaches are independent of external (world) knowledge and practice machine learning which based on corpus (Mitkov, 1999). Most of these approaches associate a score to each candidate on basis of collocation patterns, frequency or syntactic preference and combined it to select the final one antecedent with the highest value with the help of training and test corpus. For example, Nasukawa's (1994) approach used collocation patterns as selectional constraints to test the eligibility of a candidate for antecedent and implemented a heuristic rule for supporting subjects over objects. This approach also used synonym dictionary. Whereas corpus-based Dagan & Itai (1990) approach employs selectional patterns the co-occurrence patterns to resolve the references of pronoun "it" in sentences which was selected randomly by categorizing them as subject or object. On the basis of probabilistic model, the approach under this category, computes the probabilities of the antecedent from the training corpus and then uses these investigations to resolve pronouns in the test corpus (Mitkov, 1999). The number of times a referent occurs earlier in the discourse is generally considered as potential candidate for antecedent.

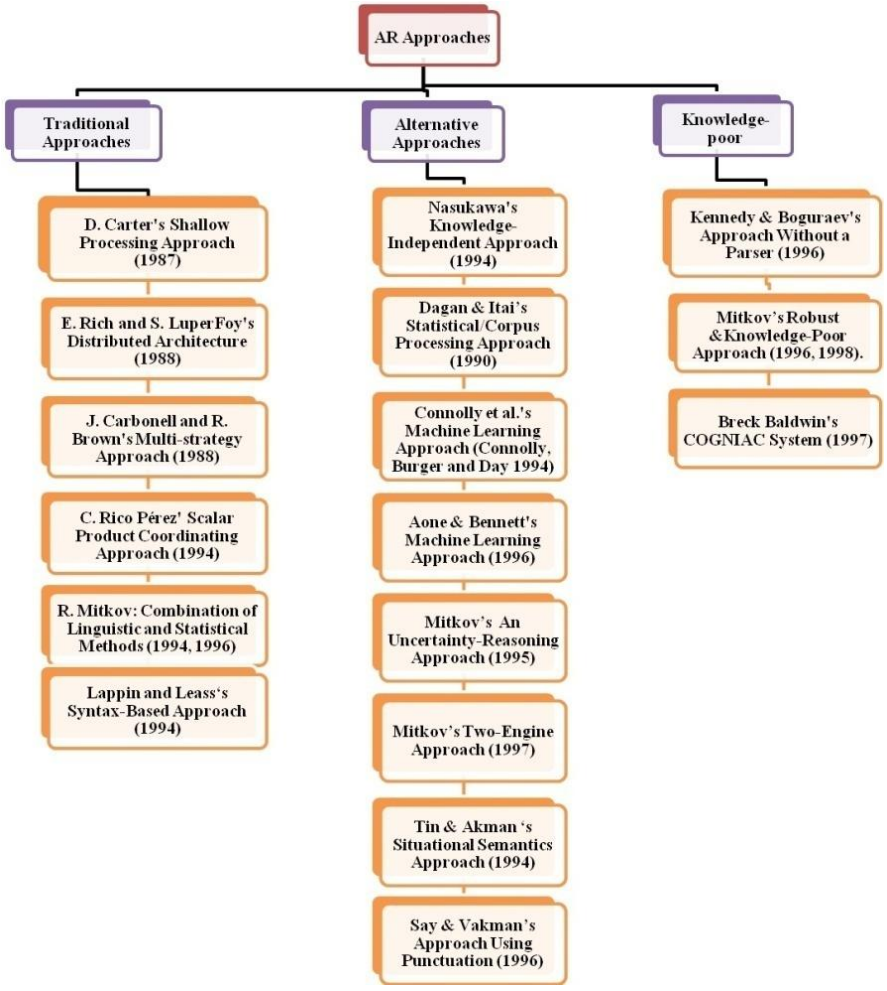


Figure 3 *Approaches in Anaphora Resolution*

3.3 Knowledge-poor approach

Knowledge-poor approach use limited methods for determining the candidates and finally selecting the potential candidate for antecedents (Mitkov, 1999). It uses very less or discard the semantic, linguistic or discourse knowledge by paying no attention to "in-depth, full" syntactic parsing for preprocessing of text. It mainly utilizes metadata of each single word/text/entity which includes POS (part of speech) tagger annotated with morphological features such as related word groups and the syntactic relations. This approach is much based on machine learning which automates almost all the preprocessing stages such as noun-phrase identification, morphological and semantic class identification, etc. Such approach either use eliminative or preferential techniques or both for finding the most suitable candidate and suggest only approximate solutions. Approaches under this category mainly use recall, precision and f-

measure as their evaluation metrics on the training and test corpora (Mitkov, 1999). It also dealt with noun phrase coreference resolution which is quite cumbersome.

The authors studied anaphora resolution for Hindi language and found that researchers in this area investigated the above discussed factors and approaches and implemented them after modification for Hindi and other Indian languages.

Sobha et al. (2000) put forward knowledge poor approach for resolving anaphora in Malayalam and Hindi with limited syntactic knowledge and salience measurement.

Agarwal et al. (2007) implemented machine learning based approach with semantic information and grammatical roles for resolving the anaphora in Hindi.

Prasad et al. (2008) used discourse based approach employing salience ranking for resolving anaphora in Hindi. Dutta et al. (2008) developed rule based approach to resolve reflexive and possessive pronouns in Hindi by employing semantic information and gender agreement.

Uppalapu et al. (2009) used discourse based approach to resolve third person pronouns in Hindi.

Dutta et al. (2011) described machine learning approach for studying indirect anaphora and classified them on basis of their semantic structure in a corpus based on Hindi language.

Chatterji et al. (2011) conducted data driven and statistical approach to determine reference of pronoun for Bengali, Tamil, and Hindi language using gender and number agreement with verb semantics.

Dakwale et al. (2013) employed hybrid approach and improved the accuracy of anaphora resolution for Hindi language by bringing in number agreement, semantic analysis and NER factors.

Lakhmani et al.(2014) used knowledge based approach and analyzed the role of Recency factor for resolving pronominal anaphora in Hindi.

Devi et al. (2014) incorporated agreement factors for analyzing the relationship between anaphors and their antecedents for Indian languages.

Mehla et al. (2015) attempted resolution of Event and Entity Anaphora for Hindi using semantic and NER information.

Mahato & Thomas (2015) implemented hybrid approach for pronominal anaphora resolution in Hindi using semantic knowledge with gender and number agreement.

4. Conclusion

It was found that the factors involved in the knowledge-rich approaches are much preference-based whereas corpus based approaches have to develop the discourse specific to the domain. On the other hand Knowledge-poor approach either none or rarely utilize domain knowledge and even do not require parsing. This literature review and research papers conclude that plenty of techniques are employed under the three computational models for anaphora resolution. The approaches discussed in the paper indicate that by utilizing the knowledge source and common set of factors with different computational strategies in an efficient and effective manner could bring out high rate of success. It is quite clear that AR process get benefited from these well known factors. The presence of gender agreement, number agreement, sentence recency, etc. definitely leads to accuracy. Researchers have used rules, corpus and/or combined both to achieve accurate results with the help supervised learning. The accuracy of the AR system much depend upon these factors and the approaches through which the factors are incorporate smartly.

References

- Agarwal, S. , Srivastava, M. , Agarwal, P., & Sanyal, R. (2007). Anaphora resolution in Hindi documents. In Proceedings of IEEE Natural Language Processing and Knowledge Engineering: 452-458.
- Chatterji, S., Dhar, A., Barik, B., Moumita PK, Sarkar, S., & Basu, A. (2011). Anaphora resolution for Bengali, Hindi and Tamil using random tree algorithm in Weka. *In Proceedings of ICON2011 NLP TOOL CONTEST: 9th International Conference on Natural Language Processing.*
- Dagan, I. & Itai, A. (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.
- Dakwale, P., Mujadia, V., & Sharma, D.M. (2013). A hybrid approach for anaphora resolution in Hindi. International Joint Conference on Natural Language Processing,: 977–981.
- Delmonte,R. , Bristot, A., Piccolino Boniforti, M.A., & Tonelli, S. (2006). Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach. Association for Computational Linguistics,11:3-10.

- Devi, S.L.,Ram, R.V., & Rao, P.R. (2014). A Generic Anaphora Resolution Engine for Indian Languages. In COLING:1824-1833.
- Dutta, K., Prakash, N., & Kaushik, S. (2011). Machine learning approach for the classification of demonstrative pronouns for indirect anaphora in Hindi news items. Prague Bulletin of Mathematical Linguistics, 95: 33-50.
- Dutta, K., Prakash, N., & Kaushik, S. (2008). Resolving Pronominal Anaphora in Hindi using Hobbs algorithm. Web Journal of Formal Computation and Cognitive Linguistics, 1(10):5607-5607.
- Jain, S., Jain, N., Tammewar, A., Bhat, R., A., & Sharma, D. M. (2013). Exploring Semantic Information in Hindi WordNet for Hindi Dependency Parsing. *The Sixth International Joint Conference on Natural Language Processing.*
- Lakhmani, P., Singh, S. (2013). Anaphora resolution in Hindi language. International Journal of Information and Computation Technology, 3(7):609-616.
- Lakhmani, P., Singh, S., & Dr. Mathur, P. (2014). Gazetteer method for resolving pronominal anaphora in Hindi language. International Journal of Advances in Computer Science and Technology, 3(3).
- Lappin S., Leass H., J. (1994). An algorithm for pronominal anaphora resolution. Computational linguistics, 20(4):535-561.
- Mahato, S., Thomas, A. (2015). Exploring Semantic Information from Hindi Dependency Treebank for Resolving Pronominal Anaphora. International Journal of Computer Applications:0975 – 8887.
- Mehla, K., Karambir, & Jangra, A. (2015). Event Anaphora Resolution in Natural Language Processing for Hindi text. International Journal of Innovative Science, Engineering & Technology, 2(1).
- Mitkov R. (1997). 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.
- Mitkov, R. (1999). Anaphora Resolution: The State Of The Art. *Proceedings of COLING'98/ ACL'98.*

- Mitkov, R, Evans, R., & Orasanal, C. (2002). A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. Lecture Notes In Computer Science, 2276:168-186.
- Moosavi, N., S., Strube, M. (2016). Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. Association for Computational Linguistics:632–642.
- Nasukawa, T. (1994). "Robust method of pronoun resolution using full-text information". *Proceedings of the 15th International Conference on Computational Linguistics(COLING'94)*, 1157-1163, Kyoto, Japan.
- Navigli, R. 2009. Word sense disambiguation: A survey. ACM Computing Surveys, 41(2).
- Nadeau, D., Sekine, S. (2006). A survey of named entity recognition and classification. *Proceedings of International Conference on Machine Learning.* National Research Council Canada / New York University.
- Prasad, R. and Strube, M. (2000). Discourse salience and pronoun resolution in Hindi. In Penn Working Papers in Linguistics, 6(3): 189-208.
- Sharma, I., Singh, P. K. (2017). A Survey on Anaphora Resolution. Recent Innovations in Computer Science and Information Technology,5-7.
- Sobha, L., Patnaik, B.N. (2000). Vasisth: An Anaphora Resolution System for Malayalam and Hindi. *Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation In Engineering and Industrial Applications.* Monastir, Tunisia.
- Surekha, S., Kumar, K. V., & Skandha, S., S. (2016). Word sense disambiguation using leak. ICRACSC:063-066
- Uppalapu, B., Sharma, D. M. (2009). Pronoun Resolution for Hindi. *In DAARC*-7.
- Yadav, D. S., Dutta, K., Singh, P. & Chandel, P. (2016). Anaphora resolution for Indian languages: The state of the art. Recent Innovations in Science and Engineering, 1(2):01-07.