ISSN 2320 - 2602

Volume 3, No.3, March 2014

International Journal of Advances in Computer Science and Technology

Available Online at http://warse.org/pdfs/2014/ijacst04332014.pdf

Gazetteer Method for Resolving Pronominal Anaphora in Hindi Language

Priya Lakhmani[1], Smita Singh[2], Dr. Pratistha Mathur[3]
[1]Banasthali Vidyapith, India, tinalakhmani@gmail.com
[2]Banasthali Vidyapith, India, smitasingh101@gmail.com
[3]Banasthali Vidyapith, India, mathurprati@yahoo.com

ABSTRACT

Anaphora processing has been an active topic of research in the field of computational linguistic. At present, resolution of anaphoric reference is one of the most challenging tasks of natural language processing domain. Most of the NLP applications such as machine translation, information extraction, question answering system, automatic summarization etc require successful resolution of anaphora. This paper focuses on pronominal anaphora resolution for Hindi Language using Gazetteer method. We have developed a model that performs pronominal anaphora resolution task for Hindi Language. There are few salient factors among which this model uses Recency factor as the baseline factor. Animistic knowledge is introduced to the model which forms the criteria of classification of different nouns and pronouns. In this paper we demonstrate three experiments conducted on different data sets containing 10 to 30 sentences in Hindi Language along with its summarized result and future directions.

Key words: Anaphora, Animistic knowledge, Gazetteer method, Natural language processing

1. INTRODUCTION

Anaphora defines an expression that is used as a reference to another expression or entities previously define in the discourse. Discourse is a group of related and collocated sentences. The process of binding the referring expression to the correct antecedent, in the given discourse, is called anaphora resolution [1].

Pronominal anaphora resolution is a subset of anaphora resolution which refers to the task of finding co-referents for pronouns. Pronominal resolution is the act of referring pronoun to the correct noun phrase. Most of the applications such as machine translation, automatic summarization, question answering system, information extraction etc requires successful identification of anaphora hence an important problem in natural language processing is the resolution of pronouns to their intended referents. This paper completely focuses on pronominal resolution for Hindi language. Consider the following sentence:

“मोहन बगीचे से अमरूद तोड़कर खाता है | उसे अमरूद पसंद है | वह फल बहु त मीठा होता है |”

This sentence demonstrates an anaphor, where the pronoun ‘उसे’ refers back to a referent. Intuitively, ‘उसे’ refers to ‘मोहन’. The pronoun ‘ वह’ refers to ‘अमरूद.’ The entity referred back to is called the ‘referent’ or ‘antecedent’. ‘वह’ is called the referring expression or ‘anaphor’; that is, the expression used to perform reference.

Anaphora resolution can be intrasentential as well as intersentential. Intrasentential is the case where the antecedent is in the same sentence as that of anaphor. Consider the sentence,

“मोहन ने सीता को उसकी पुस्तक दी |”

Here ‘उसकी’ refers to ‘सीता’. This is an example of intrasentential anaphora, whereas intersentential refers to antecedents that are in a different sentence to the anaphor. In these sentences,

“राम एक किसान का पुत्र है |
उसे खेतों में काम करना पसंद है |”

‘उसे’ refers to ‘राम’. This is an example of intersentential anaphora. When performing anaphora resolution all noun phrases are typically treated as potential candidates for antecedents. The scope is usually limited to the current and preceding sentences and all candidate antecedents within that scope are considered.

Every language has its own structure and grammar. In Hindi language pronoun exhibits great deal of ambiguity. A pronoun in Hindi does not provide any information about gender. In Hindi language there is no differentiation between ‘he’ and ‘she’. ‘वह’ is used for both the gender and is decided by the verb form. For number marking, in Hindi, some forms, like ‘उसको’(him), ‘उसने’(he) are unambiguously singular but some forms can be both singular and plural, like ‘उन्होने’ (he)(honorific)/they, or ‘उनको’(him)(honorific)/ them. So resolving pronoun in Hindi is complex task to be handled.

2. RELATED WORK

An extensive research work for anaphora resolution is majorly classified by three main algorithm developed by researchers.

- First work in the field of pronoun resolution is done by J.R Hobbs in English language in 1976. Hobb's algorithm makes use of syntactic information for resolving pronoun. It gave accuracy of 82% for English language [5].
- Joshi, A. K. & Kuhn. S, in 1979 and Joshi, A. K. & Weinstein.S in 1981, gave centering theory for pronoun resolution. This work is also done in English language which gave 76% accuracy [6].
- S. Lappin and H. Leass proposed their algorithm for pronoun resolution for English language in year 1994. Lappin and Leass evaluated RAP (*Resolution for anaphora procedure*) using 360 pronoun finds the correct antecedent for 310 pronouns, 86% of the total (*74% of intersentential cases and 89% of intrasentential cases*)[7].

The work done for anaphora resolution based on Gazetteer method is summarized below:

- Richard Evans and Constantin Orasan improved anaphora resolution by identifying animate entities in texts [4].
- Ruslan Mitkov, Richard Evans resolved anaphora resolution by using Gazetteer method in 2007[2].
- Tyne Liang and Dian-Song Wu used above approach in automatic pronominal anaphora resolution in English texts in 2002.
- Constantin Orasan and Richard Evans used NP Animacy Identification for Anaphora Resolution in 2007[2].
- Natalia N. Modjeska, Katja Markert and Malvina Nissim used web in Machine Learning for Other-Anaphora Resolution in 2003[3].

3. SALIENT FACTORS

Resolving pronoun for Hindi language requires various factors to be considered. The following factors play an important role for pronominal resolution in Hindi language.

- ***Recency:*** A proposal source, Recency moves backwards spatially through the text and adds noun phrases to the blackboard as candidates. The confidence score is set on proposal as a float value starting at one and exponentially decreasing to zero as the proposer reaches the beginning of the analyzed text.
- ***Gender Agreement:*** Gender Agreement compares the gender of candidate co referents to the gender required by the pronoun being resolved. Any candidate that doesn't match the required gender of the pronoun is removed from further consideration.
- ***Number Agreement*:** Number Agreement extracts the part of speech of candidates. The part of speech label is checked for plurality. If the candidate is plural but the current pronoun being resolved doesn't indicate a plural co referent, the candidate is removed from consideration. The same process occurs for singular candidates which are removed if the pronoun being resolved requires a plural co referent.
- ***Animistic Knowledge*:** Animistic knowledge filters candidates based on which ones represent living beings. Inanimate candidates are removed from consideration when the pronoun being resolved must refer to an animated co referent, and animated candidates are removed from consideration for pronouns that must refer to inanimate co referents.

Contribution of these factors to anaphora resolution increases accuracy of resolving system.

4. APPROACH

Classification of Anaphora Resolution method is mainly done by three main approaches- Syntax based approach (*Hobbs algorithm*), Discourse based approach (*Centering algorithm*) and Hybrid approach (*Lappin Leass algorithm*). We have developed a computational model which uses Gazetteer method also called list look up method for pronominal resolution in Hindi language.

4.1 Gazetteer Method

Gazetteer Method is the method which creates different gazetteer classes (*lists*) for different elements and then applies operations to classify the elements. Gazetteers are utilized to supply external knowledge to learners, or to supply annotated data with a training source. In our system we have created lists of animistic pronoun, animistic noun, non animistic pronoun, non animistic noun and middle animistic pronoun. The elements are classified into list according to their property. That is why, this method is also called list look up method. This external knowledge helps the system in resolving anaphora by differentiating the co referents on the basis of their classification.

4.2 Advantages of Gazetteer Method

- The Gazetteer method provides very fast result of Anaphora Resolution System
- The accuracy of Gazetteer method depends on completeness of the Gazetteer used.
- Gazetteer method increases the system's accuracy to far extent.

5. RESOLVING SYSTEM

A computational model for resolving anaphora has been developed based on the above mentioned factors in which Recency factor and animistic knowledge have their significant contribution. The resolving system uses Recency as a baseline factor for resolving anaphora. Animistic knowledge is used for learning the system, and guides the system to differentiate between animate and inanimate things.

5.1 Working of system

Our system first finds out the referent for pronoun using Recency factor. Recency factor describes that the referents mentioned in current sentence tends to have higher weights than those in previous sentence. Recency factor assigns the highest weight for a pronoun co referent to the first previous noun detected while parsing backward. For example, consider the sentence

“सुनीता ने गीता को गुलाब दिया|
वह बहु तखूबसूरत था|”

In this sentence there are three nouns ‘सुनीता’, ‘गीता’, ‘गुलाब’. According to Recency factor the highest weight is assigned to the closest noun ‘गुलाब’ for the pronoun ‘वह’. Our system uses a concept from Lappin Leass approach [7] is used for finding referent using Recency as a salient factor.

To increase the accuracy of the system we use animistic knowledge by training the data especially nouns and pronouns. We have created different classes for animistic nouns (*nouns which are living beings*), animistic pronouns (*pronouns which only refers to living beings*), non animistic noun (*non living nouns*), non animistic pronoun (*which only refers to non living things*) and last category is middle animistic pronoun (*pronoun which refers to both living and non living objects*). Consider the following example:

“सीता फल खाती है और अपने बच्चों को भी खिलाती है|”

Here ‘अपने’ is an animistic pronoun which can only refer to animistic noun ‘सीता’. ‘अपने’ cannot refer to ‘फल’ because ‘फल’ is a non animistic noun. The resolving system performs the task of resolution in following manner:

1. When the system encounters any pronoun then first it finds the referent noun based on Recency factor. Hence it chooses the closest noun as a referent.
2. The system checks whether the pronoun falls under animistic, non animistic or middle animistic category.
3. If the pronoun falls under animistic category then it checks whether the referent selected by Recency factor falls under animistic noun or non animistic noun category.
4. If the referent selected falls under animistic noun category then that referent is the final output for that pronoun otherwise if the referent falls under non animistic noun then in that case the referents are backtracked (*at least up to three sentences*) until we find the correct animistic referent for animistic pronoun.
5. If the pronoun falls under non animistic category, then the same process mention above is done until we get a non animistic referent.
6. If the pronoun falls under middle animistic category then the referent selected by Recency factor is the final output.

6. EXPERIMENTS AND RESULTS

A standard experiment is based on finding the contribution of Recency factor and animistic knowledge to the overall accuracy of correctly resolved pronouns. Experiments are performed on the different data set in order to identify the overall success of the system. The correctness of the accuracy obtained by the experiments is measured by the language expert.

6.1 Data Set 1

This experiment uses the text from children story domain. We have taken short stories in Hindi language from indif.com(*http://indif.com/kids/hindi_stories/short_stories.aspx*), a popular site for short Hindi stories and performed anaphora resolution over these stories. Ideally this experiment represents a baseline performance since the story is a straightforward narrative style with extremely low sentence structure complexity. Also it contains approx 10 to 25 sentences having 100 to 300 words. The result shown by experiment is summarized in table 1:

Table 1: Result from experiment performed on short stories

Data Set	Total Sentences	Total Word	Total Anaphor	Correct Resolved Anaphor	Accuracy
Story1	11	129	13	10	77%
Story2	11	133	11	9	82%
Story3	23	275	22	7	32%
Story4	17	213	20	12	60%
Story5	21	227	21	11	53%

The result of the experiment on short story shows that the resolving system is 61% accurate on an average.

6.2 Data Set 2

This experiment used the text from a news article. We have taken news articles in Hindi language from webduniya.com (*http://hindi.webdunia.com/news*), a popular site for Hindi news and perform anaphora resolution task over these articles. We have taken five news articles of different genres which include sports, entertainment, science and general topics. The accuracy shown by experiment is summarized in table 2:

Table 2: Result from experiment performed on news articles

Data Set	**Total Sentences**	**Total Word**	**Total Anaphors**	**Correctly Resolved Anaphor**	**Accuracy**
News1	9	175	7	3	43%
News2	8	207	6	3	50%
News3	8	143	10	6	60%
News4	13	247	18	15	83%
News5	11	195	14	10	72%

The result of the experiment shows that Recency factor and animistic knowledge provides approx 62% success to overall system. It is observed that success rate vary with the order of words and the type of sentences.

6.3 Data Set 3

This experiment uses text from Wikipedia in Hindi language (*http://hn.wiki.org*). We have taken biography of famous political leaders of India from this site. The following result is obtained is summarized in table 3:

Table 3: Result from experiment performed on biography

Data Set	**Total Sentences**	**Total Word**	**Total Anaphors**	**Correctly Resolved Anaphor**	**Accuracy**
Wiki1	16	329	16	13	82%
Wiki2	20	347	16	14	88%
Wiki3	22	374	17	12	71%
Wiki4	14	282	12	10	84%
Wiki5	28	348	19	15	79%

The result of above experiments demonstrates that resolving system shows approx 82% success rate. It is observed that success rate vary with the structure of sentences. As Hindi is a free word, the success depends on the writing style of the text. Different articles have different way of writing. This affects the accuracy of the system.

From the above experiments, articles about the political leaders from Wikipedia show the highest success rate. Hence success of the resolving system depends on the type of input text document given to the system. By using Recency factor and animistic knowledge our system is approximately giving 60% to 70% accuracy. More factors can be added such as gender agreement and number agreement in order to increase the accuracy of overall system.

7. CONCLUSION

This paper presents a computational model based on gazetteer method for resolving anaphora in Hindi Language. A standard experiment is performed on different data set having different style of written text in Hindi Language each having 100 to 300 words. The experiment is conducted taking Recency as baseline factor and animistic knowledge is induced to the system for learning the system to differentiate between animate and inanimate nouns and pronouns. The system gives approximate 60 to 70 percentage successful identification of anaphora. For future work we can add other factors such as gender agreement and number agreement in order to increase the success rate of our resolving system. Also, this work has not been done in other Indian languages. So there is a large scope of anaphora resolution to be done in other Indian languages.

REFERENCES

1. Denber M. **Automatic Resolution of Anaphora in English**, *Technical Report, Eastman Kodak Co. Imaging Science Division*, June 30, 1998.
2. Constantin Orasan and Richard Evans. **NP Animacy Identification for Anaphora Resolution**, *Journal of Artificial Intelligence Research* 29 (2007) 79-103
3. Razvan Bunescu. **Associative anaphora resolution: A web-based approach.** *In Proceedings of EACL* 2003 - *Workshop on The Computational Treatment of Anaphora*, Budapest 2003.
4. Barlow M. **Feature Mismatches and Anaphora Resolution.** *In Proceedings of DAARC2, University of Lancaster.* 1998
5. Hobbs, Jerry. **Resolving Pronoun References** in *B. Grosz. K. Sparck ' Jones, and B. Webber (eds.) Readings in Natural Language Processing.* California: Morgan Kaufman Publishers Inc 1986
6. Joshi, A. K. & Kuhn. S; Joshi,A. K. & Weinstein.S **Centering theory based approaches** 1979, 1981
7. Thiago Thome, Coelho. **Lappin and Leass algorithm for pronoun resolution in Portuguese**, *EPIA'05 Proceedings of the 12th Portugues conference on Progress in Artificial Intelligence* , Brazil, Pages 680-692
8. Ruslan Mitkov, Richard Evans **Anaphora Resolution: To What Extent Does It Help NLP Applications?**, *Springer-Verlag Berlin Heidelberg 2007, DAARC 2007, LNAI 4410*, pp. 179–190, 2007
9. Brent, **from grammar to lexicon: unsupervised learning of lexical syntax**. *Computational Linguistics*, 19(3):243–262. 1993
10. Munoz, R., Saiz-Noeda, M., Montoya, A. **Semantic information in anaphora resolution.** *In Proceedings of Portal*, 63-70, 2002.