# PRONOMINAL ANAPHORA RESOLUTION IN PUNJABI LANGUAGE

Priya Lakhmani[1], Smita Singh[2], Dr. Pratistha Mathur[3], Dr. Sudha Morwal[4]

Department of Computer Science, Banasthali University, Jaipur, India

***ABSTRACT***

*Anaphora Resolution is a process of finding referents in discourse. In computational linguistic, Anaphora resolution is complex and challenging task. This paper focuses on pronominal anaphora resolution. It is a subpart of anaphora resolution where pronouns are referred to noun referents. Including anaphora resolution into many applications like automatic summarization, opinion mining, machine translation, question answering systems etc. increase their accuracy by 10%. Related work in this field has been done in many languages. This paper focuses on resolving anaphora for Punjabi language. A model is proposed for resolving anaphora and an experiment is conducted to measure the accuracy of the system. The model uses two factors: Recency and Animistic knowledge. Recency factor works on the concept of Lappin Leass approach and for introducing animistic knowledge gazetteer method is used. The experiment is conducted on a Punjabi story containing more than 1000 words and result is drawn with the future directions.*

***KEYWORDS***

*Anaphora, Discourse, Lappin Leass approach, Gazetteer method, Natural Language Processing*

## 1. INTRODUCTION

Anaphora is a process of referring back to previous element in the discourse. Discourse is a group of collocated and inter related sentences. Anaphora Resolution is defined as the problem of identifying referents in the discourse. Consider the following:

"**Arunima** went to market and bought a dress.
**She** gave it to Deepa."

This is an example of anaphora resolution. Here "She" is an anaphora which refers to "Arunima". The entity which is referred back is called either 'referent' or 'antecedent'. Here "Arunima" is antecedent.

This paper completely focuses on pronominal anaphora resolution. It's the most common type of anaphora. Pronominal anaphora resolution is the process of finding noun phrase which refers to pronoun and it occurs at the level of personal pronoun, possessive pronoun, demonstrative pronoun, reflexive pronoun and relative pronouns.

Though anaphora resolution task seems very simple it can become increasingly complex when we encounter sentences like:

"Fruits were given to children because they were there."

In the above sentence "they" is either referred to "fruits" or "children". This anaphor creates ambiguity & resolves to either or both. Hence, this requires semantic and pragmatic knowledge for performing anaphora resolution task. Figuring out what expressions in a text refer to same

DOI:10.5121/ijfcst.2014.4408 

entity enables a system to correctly binding facts to the appropriate internal representations of the entities that have been recognized. Therefore anaphora resolution is one of the active research areas within the realm of Natural Language Processing (NLP).

## 2. RELATED WORK

The Extensive work is done in the field of anaphora resolution for Indian and European languages. A short summarization of this work is:

- 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[1].
- Tyne Liang and Dian-Song Wu used above approach in automatic pronominal anaphora resolution in English texts in 2002[16].
- 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].
- Anaphora resolution system for German language based on extension of Centering theory was presented by Strube & Hahn in 1991[6].
- An algorithm for pronoun resolution for English language was proposed by S. Lappin and H. Leass in year 1994[15].
- Joshi, A. K. & Kuhn. S, in 1979 and Joshi, A. K. & Weinstein.S in 1981, presented a new theory called centering theory for pronoun resolution [8].
- Pronominal anaphora is also resolved in Nepali Language using Lappin Leass approach by Dev Bahadur [9].
- Thiago Thomes Coelho, Ariadne Maria Brito Rizzoni done work in Portugeese language using Lappin and Leass algorithm [7].
- Anaphora resolution is also done in Spanish Texts using Centering approach by Manuel Palomar, Lidia Moreno and Jesfis Peral [10].
- S.Lappin and M.McCord developed a syntactic filter on pronominal anaphora for slot grammer using Lappin Leass principles in 1990[11].
- Sobha and Patnaik presented a rule based approach for the anaphora resolution in Hindi language and Malayalam language [12].
- Dutta et al. presented modified Hobbs algorithm for Hindi [13].
- J.Balaji applied Centering principles in Tamil [14].

## 3. CHALLENGES

There are certain issues which are needed to be considered while performing anaphora resolution in Punjabi language. These are mentioned below:

- *Encoding in standard form:* Large amount of information is available in Punjabi on www (on electronic document form). But there is no standard form i.e. information is encoded in different fonts. Hence it becomes difficult for implementation.
- *Requirement of Unicode based tools for Punjabi:* Unicode based font are very problematic as Unicode based tools may not support Punjabi language. Hence, due to lack of standardization it becomes difficult to use these documents in developing corpus.
- *No Capitalization:* Concept of Capitalization is not present in Punjabi Language.

- *Morphological and inflectionally rich:* Punjabi is morphological and inflectionally rich language. Also, it is a free word order. There is no fixed order of subject, object, and indirect object. This causes difficulty in resolving pronouns.

## 4. SALIENT FACTORS

The model proposed for anaphora resolution uses Recency factor and Animistic knowledge for resolving pronominal anaphora in Punjabi language.

### 4.1. Recency

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,

"ਸੀਤਾ ਇੱਕ ਗੁਲਾਬ ਦੇ ਖਰੀਦਿਆ | ਇਹ ਸੁੰਦਰ ਹੈ |"

In this sentence there are two nouns "ਸੀਤਾ" and "ਗੁਲਾਬ". Recency factor assigns the highest weight to the closest noun "ਗੁਲਾਬ". Hence, the pronoun "ਇਹ" refers to "ਗੁਲਾਬ". Most of the times Recency factor gives correct resolution but sometimes it fails to identify correct referent. So, animistic knowledge is added for successful identification of anaphora.

### 4.2. Animistic Knowledge

Animistic knowledge is introduced to the system in order to differentiate between living and non living entities. Animate entities include people and animals. Animate pronouns should refer to animate nouns. Inanimate co referents are eliminated from consideration when the pronoun being resolved is an animate pronoun, and animate co referents are eliminated from consideration for non animistic pronouns that must refer to inanimate co referents. Consider the following:

"ਨੇਹਾ ਨੇ ਆਪਣੇ ਲਈ ਇੱਕ ਪੈਨ ਖਰੀਦਿਆ"

In the above example pronoun "ਆਪਣੇ" is animistic pronoun (*always refer to living things*). So, it refers to animistic noun "ਨੇਹਾ".

In addition to Recency and Animistic factor, there are two more factors that affect the anaphora resolution. These are gender agreement and number agreement.

### 4.3. Gender Agreement

Gender Agreement matches the gender of co referents with the gender of the pronoun which is to be resolved. The co referent that doesn't suits with the pronoun in terms of male and female is eliminated from further consideration.

### 4.4. Number Agreement

Number Agreement checks for plurality. Singular pronoun should refer to singular co referent and plural pronoun should refer to plural co referent. If the co referent is plural but the pronoun being resolved is singular then the co referent is eliminated from consideration and vice versa. For example,

"ਸੀਤਾ ਅਤੇ ਗੀਤਾ ਦੋਸਤ ਹੁੰਦੇ ਹਨ | ਉਹ ਖੇਡਣ ਲਈ ਚਾਹੁੰਦੇ|"

In the above example, "ਉਹ" refer to "ਸੀਤਾ ਅਤੇ ਗੀਤਾ".

# 5. ANAPHORA RESOLUTION SYSTEM

## 5.1. Lappin Leass Approach

The system uses Lappin and Leass approach for applying Recency factor. This approach falls under the category of hybrid approach. This approach is based on the fact that pronouns are more likely to refer to entities mentioned recently in the discourse. The algorithm involves calculating salience values for each new entity that is encountered in a noun phrase. These salience values are calculated by summing the weights assigned to various factors. [15].

## 5.2. Gazetteer Method

This method is used to provide animistic knowledge to the system. In this method lists are created. These lists also called classes or Gazetteers. Elements present in the list are then classified based on certain operations. Therefore it is also called List Look Up method.

In the proposed model lists are created for nouns and pronouns based on animistic factor. List for animistic pronoun (*pronoun refer to living things*), non animistic pronoun (*pronouns refer to non living things*), middle animistic pronoun (*pronouns refer to both living and non living things*) are created. Lists of animistic noun (*always represent living things*) and non animistic noun (*always represent non living things*) are also created.

## 5.3. Working of the system

The system first classifies all the nouns and pronouns extracted from the input documents. Then it finds out the referent or antecedent for referencing expression based on Recency factor and store it as intermediate result. The previous closest noun is chosen as a referent for the anaphora. This antecedent is then verified from the list based on animistic knowledge in order to find correct referent for the anaphora and then final output is displayed. 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.

The following flowchart shows the working of overall system for anaphora resolution:

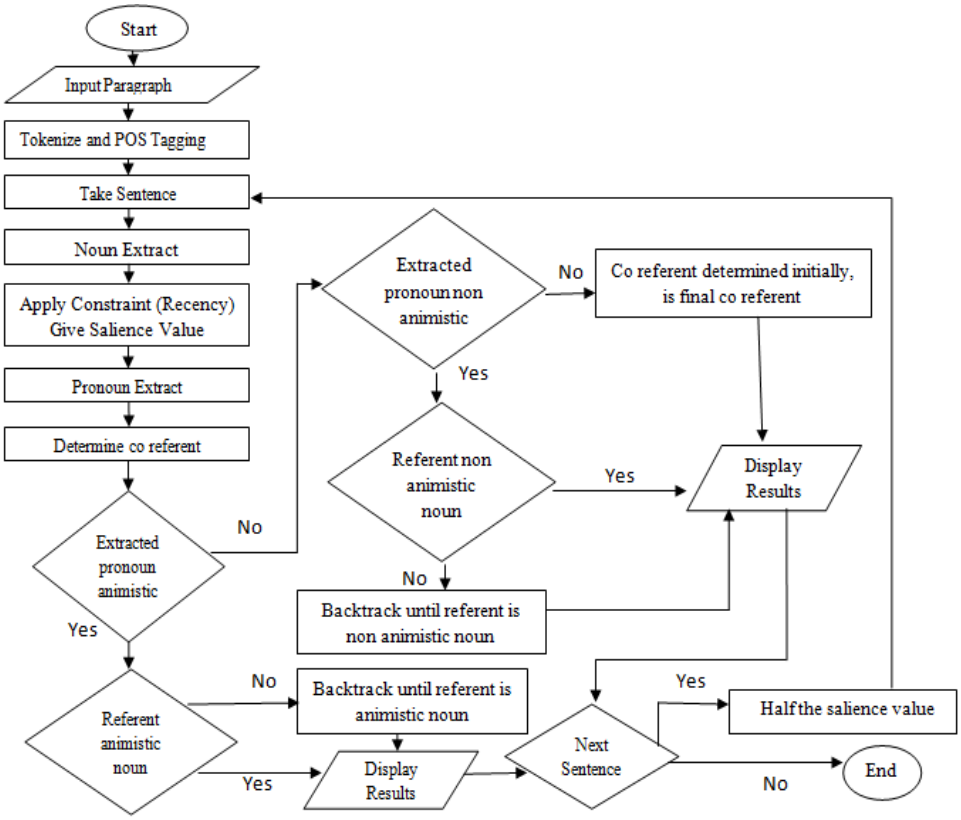


Figure 1. Flowchart of the system

## 6. EXPERIMENT AND RESULT

A standard experiment is based on finding the contribution of Recency factor and Animistic knowledge to the overall accuracy of correctly resolved pronouns. Recency factor is taken as a baseline factor. Then animistic knowledge is added to increase the accuracy of the overall system.

### 6.1. Data Set

The experiment uses the text from story domain. We have taken long story in Punjabi language from (https://sites.google.com/site/punjabisahit/home/punjabi-stories/marichika-maricika) a popular site for Punjabi stories and performed anaphora resolution over the POS tagged story. The story is a straightforward narrative style with extremely high sentence structure complexity. The result of experiment is summarized in Table 1:

Table 1. Result of experiment

| | |
|---|---|
| Total No of Sentences | 75 |
| Total No of words | 1341 |
| Total No of Anaphors | 117 |

| | |
|---|---|
| Correctly resolved anaphora by Recency factor only | 35 |
| Correctly resolved anaphora by Recency factor and Animistic knowledge | 74 |

The correctness of the accuracy obtained by the experiment is measured by the language expert. The result of this experiment shows that Recency provides approx 30% accuracy which proves that Recency factor alone cannot resolve pronoun correctly, some more factors should be added. Adding animistic knowledge to the system increases the accuracy to 64%. Still there are some pronouns which are not resolved correctly. More factors can be added such as gender agreement, number agreement, and pragmatic knowledge in order to increase the accuracy of overall system.

## 7. CONCLUSION

This paper proposes a model for anaphora resolution task in Punjabi language. The model uses Recency factor as a baseline factor. Animistic knowledge is induced in order to increase the accuracy of the system. Gazetteer method is used for introducing animistic knowledge to the system. An experiment is performed on a Punjabi story containing more than 1000 words. The result gives 64% success to the overall system. Remaining pronouns can be resolved correctly by adding semantics and pragmatic knowledge to the system. Since Punjabi is morphological and inflectionally rich. Also it is free word order. This affects the structure of sentences and hence affects the accuracy. Also we considered only two factors (*Recency and Animistic knowledge*). There are other factors like Number Agreement and Gender Agreement that affect the accuracy. In the future we will try to incorporate these factors in our system in order to increase the success rate of the system. Also, some more experiments will be conducted on different genres of texts in order to calculate the overall accuracy of the resolving system.

## REFERENCES


[1] Ruslan Mitkov, Richard Evans, (2007) "Anaphora Resolution: To What Extent Does It Help NLP Applications?" *DAARC*, LNAI 4410, pp. 179–190.

[2] Constantin Orasan and Richard Evans ;( 2007) "NP Animacy Identification for Anaphora Resolution", Journal of Artificial Intelligence Research 29, 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., (1998). Feature Mismatches and Anaphora Resolution. In Proceedings of DAARC2, University of Lancaster.

[5] Brent, (1993). "From grammar to lexicon: unsupervised learning of lexical syntax". Computational Linguistics, 19(3):243–262.

[6] Strube & Hahn "A system for anaphora resolution for German based on extension of Centering theory".

[7] Thiago Thomes, "Lappin and leass algorithm for pronoun resolution in Portuguese", Institute of State University of Campinas, Campinas, SP, Brazil EPIA'05 Proceedings of the 12th Portuguese conference on Progress in Artificial Intelligence Pages 680-692.

[8] Aravind K Joshi, Rashmi Prasad, and Eleni Miltsakaki "Anaphora Resolution: A Centering Approach".

[9] Dev Bahadur Poudel and Bivod Aale Magar "Anaphoric Resolution in Nepali", Nepal Engineering College.

[10] Manuel Palomar, Lidia Moreno "Algorithm for Anaphora Resolution in Spanish Texts", University of Alicante, Valencia University of Technology.

[11] McCord, Michael, (1990)"Slot grammar: A system for simpler construction of practical natural language grammars." In Natural Language and Logic: International Scientific Symposium, edited by R. Studer, 118-145. Lecture Notes in Computer.
[12] L. Sobha and B.N. Patnaik, “Vasisth: An anaphora resolution system for Malayalam and Hindi”, Symposium on Translation Support Systems, 2002.
[13] K. Dutta, N. Prakash and S. Kaushik, “Resolving Pronominal Anaphora in Hindi using Hobbs algorithm,” Web Journal of Formal Computation and Cognitive Linguistics, Issue 10, 2008.
[14] Anaphora Resolution in Tamil using Universal Networking Language "12/2011; In proceeding of: Indian International Conference on Artificial Intelligence (IICAI-2011), At Tumkur, Karnataka, India.
[15] Shalom Lappin and H.J. Leass. 1994. “An algorithm for pronominal anaphora resolution.” Computational Linguistics, 20(4):535 – 562.
[16] Tyne Liang and Dian-Song Wu. (2004) “Automatic Pronominal Anaphora Resolution In English Texts” Computational Linguistic and Chinese Language Processing, Vol 9. No.1: 21-40.