ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/100194, August 2016*

Improving Anaphora Resolution by Resolving Gender and Number Agreement in Hindi Language using Rule based Approach

A. Ashima and B. Rajni Mohana*

Jaypee University of Information Technology, Waknaghat, P.O. Waknaghat Teh Kandaghat, Distt Solan - 173 234,
Himachal Pradesh, India;
ashi.chd92@gmail.com , rajni.mohana@juit.ac.in

Abstract

Anaphora resolution has been an active research area in the field of natural language processing. There are various methods to resolve the pronouns that refer to nouns but this paper presents the rule based approach to resolve the anaphora. The rules have been derived by the execution of the instruction from a set of data which consist of a specific type of knowledge of verbs in Hindi language. The Hindi language is resource poor language which has no specific structure of writing. We have proposed a model that performs pronominal anaphora resolution task based on gender and number agreement. Anaphora resolution is very complex and challenging task to handle so we have analyzed the similarities and variations between their agreement and pronouns in Hindi language. This paper exhibits the experiments that are governed on different data sets in Hindi Language along with its result based on the F-score and future directions.

Keywords: 5-6 words, Drawn from title, Word representing the work

1. Introduction

Indian language has variety of anaphoric utterance and these utterances bring grace and make the Indian language text interesting to read. Anaphora utterance in a document refers to another item in a document. The process of solving the anaphors is called anaphora resolution so it is the task of finding pronoun that refer to noun phase in discourse.

Eg. :- अंकुर पुस्तकालय के लिए चला जाता है और वहां से वह एक कतिाब जारी करता हैं।

In this example "वहां" refers to "पुस्तकालय" and "वह" refers to "अंकुर" it is easily understandable to human . It is tedious task to implement in software. This makes anaphora resolution as one of the most challenging tasks in the field of Natural Language Processing (NLP). Anaphora resolution is required in NLP applications like Machine translation, automatic summarization etc. to obtain the good results by solving the pronouns.

Many researchers have worked in this domain but they have focused on English and not so great work has been done in Indian parochial languages that are Hindi, Bengali, Malayalam, Tamil and Punjabi etc. We are targeting on Anaphora resolution in Hindi. Anaphora Resolution in Hindi is a hard-won as well as confusing job as it is a self- ruling language which is independent from the word order. In Hindi the anaphora resolution has many issues like Recency factor, Named entity recognition, Animistic Knowledge, Gender, Number agreement etc.

2. Gender Agreement

It differentiates the gender of person whether the gender is masculine or feminine with reference to the gender

*Author for correspondence

specification that is appropriate to the pronoun which is being resolved.

E.g.: प्‌रभात हलवाई की दुकान के पास गया और वहां से वह मठिाई खरीदता है।

कनकिा हलवाई की दुकान के पास गई और वहां से वह मठिाई खरीदती है।

The verbs are used to resolve pronouns based on gender agreement in Hindi language. In the

Above example from the verbs "खरीदता है" and "खरीदती है", it can be easily understand that "वह" refers to masculine and feminine respectively.

3. Number Agreement

It takes out part of speech of items and checked for singularity and plurality.

E.g.: रषिभ और शैल भाई हैं और वे छुट्‌टयिों के लिए गोवा के लिए चले गए हैं।

In this example the pronoun "वे" point to "रषिभ और शैल" that is plural.

In this paper we are tried to resolve the gender and number agreement by using rule base approach. Rule based approach combines knowledge sources and various factors which removes the items that are not required in a list until a set of plausible item is obtained. The constraints (rules) work as a filter to remove the unwanted item within a set of defined rules.

Hindi has no proper semantic and syntactic structure like English. The expression is changed with every sentence which causes the great deal of confusion. The paper is divided into 4 sections Section 2 gives a brief history of the previous works on the Anaphora resolution. Section 3 presents the proposed approach and 3.1 presents the flowchart of system.

Section 4 exhibits the experiments and results. At last, the paper concludes in Section 5.

4. Related Work

An extensive research has been done in various Foreign Languages like English, Arabic, Chinese etc. The experimenters have also worked on the anaphora resolution in Indian territorial Languages like Hindi, Punjabi, Bengali, Malayalam and Tamil etc. From the past ten years , a great number of models based on different approaches to resolved anaphora resolution in Hindi domain have been designed. A work done in Anaphora resolution in Hindi domain summated below:

In[1] designed a model of natural language interface (NLI) by using databases for Hindi language for resolving Reflexive Pronoun, Possessive Pronoun, and Demonstrative Pronoun but not resolved the pronouns based on gender. In[2] resolved the problem of syntactic and semantic structure of the Hindi language in Pronominal resolution and resolved all constraint resources by using knowledge based approach but gender agreement had no contribution to performance. In[3] resolved the pronominal resolution by using Gazetteer method and covered the Animistic and Recency factor but did not cover the Number and Gender agreement. In[4] handled the unknown words in Named Entity Recognition by using Transliteration approach in Hindi. In[5] used machine learning approach for the classification of indirect anaphora in Hindi text and based on the semantic structure. In[6] gave a generic anaphora engine for Indian languages and used and used CRFs, a linear graphical machine learning algorithm to train the system to resolved gender issue. In[7] presented a hybrid approach with dependency structures and a rule-based module to resolve Entity-pronoun references in Hindi. In[8] gave an improved S-List algorithm to resolve the Hindi third person pronouns. In[9] proposed a model by using rule-based translation methodology to resolve anaphora resolution English to Hindi. In[10] modified the Hobb's algorithm into the Hobb's naïve algorithm for Hindi language to resolve the anaphora resolution.

It was observed that less work done taking verbs to the text into consideration. So we used verb to resolve the both gender and number agreement. The F-score of the overall system is 79%.

5. Proposed Approach

Hindi language is morphologically rich and independent from the word order language which illustrate a large confusion. We do not gain any knowledge about the gender from the pronoun like first, second and third for example "वह " is the third person singular pronoun that is used for both male and female .Other pronouns , like 'उसने ', 'उसको' are used for both male and female and these are masculine and feminine singular but few pronouns can be both masculine ,feminine singular and plural, like 'उन्‌होंने ', 'उनको '. The proposed system consists of rules framed by identifying the various structure of Hindi language. This system extracts the verb from the input sentence and applies the rules to find out the num-

Rule1:-If the gender of the verb contains ‘a’/‘ ‘ at the last then the noun having a masculine gender.
Rule2:-If the gender of the verb contains ‘e’/‘ ‘and‘ ’ at the last then the noun having a feminine gender.
Rule3:-If the gender of the verb contains ‘a’/‘ ‘ at the last then the pronoun having a masculine gender.
Rule4:-If the gender of the verb contains ‘e’/‘ ‘and‘ ’ at the last then the pronoun having a feminine gender.
Rule5:-If the word of the verb contains ‘ae’/‘ ‘and‘aae’/‘ ’ at the last then the noun /pronoun is plural.
Rule6:-If the word of the verb contains ‘au’/“ “,‘e’/‘ ‘ and‘a’;‘ ’at the last then the noun/pronoun is singular.

Figure 1. Rules for gender and number agreement.

ber and gender of the pronoun. It also maps the pronoun to the noun. The defined rules are given below in [Figure 1]:-

For example In Rule 1:- If the gender of the verb contains ‘a’/ ‘ा ‘at the last then the noun having a masculine gender. For example :” राहुल बाज़ार जा रहा था |” Now in this the verb” रहा” ends with “ ” that means Rahul is a masculine gender. In Rule 2:- If the gender of the verb contains ‘e’/ ी ‘ and ‘ई’ at the last then the noun having a feminine gender. “मीना घर चली गयी |” Now in this verb “चली “ ends with ी ” that means Meena is a feminine gender. Similarly it applies to other rules.

5. Flowchart of the System

The working of the proposed system is shown in [Figure 2]:-

1. The data is taken as input than it is chopped into pieces the process is known as tokenization.
2. Then the tokens are added into the list where it will be trained with word net library so that we can calculate part of speech (POS) tagging which marks the word in text corresponding to the particular part of speech.
3. Than the token is extracted from the list to check if it is a verb than it will be added to the list of verb and if not than it is a pronoun than it will be added to pronoun list.
4. Now the token is checked if it is not pronoun than system will extricate newly token from the list.
5. Now the token is again checked if it is a first pronoun than the verb is assigned to that token. After that the gender agreement rule and then number agreement rule is applied to resolve the pronoun based on gender and number information.
6. If it is not the first pronoun than it will again extract the token from the list.
7. If all the pronouns are resolved than it will end the task and if not than it will again extract the tokens from the list.

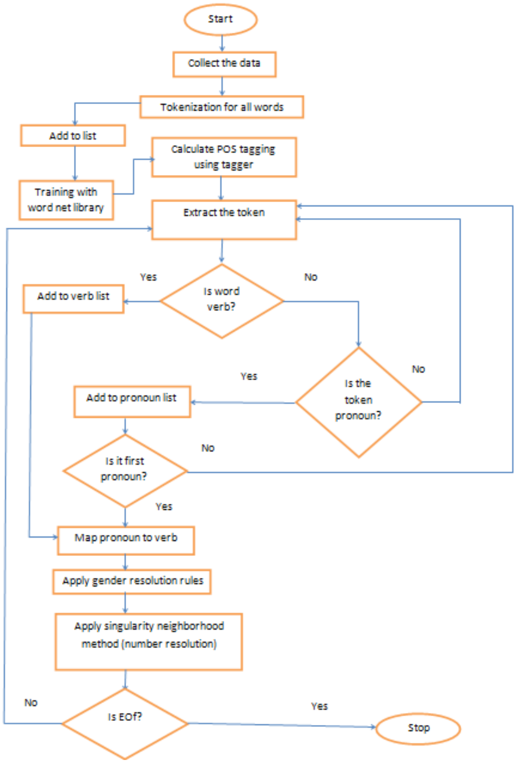

Figure 2. Flowchart of the System.

6. Experiment and Result

We have tested our approach on two different types of data sets. The rule based approach is based on finding the gender and number agreement. Based on gender and number agreement the F-score of the system is calculated. The F- score combines precision and recall[11] The formulas are presented below:

RECALL is defined as the number of correct entities retrieved from the document is divided to the total number of correct entities in the document. It showed as a percentage.

Recall:R=

$$\frac{\textit{Correct entities detected by system}}{\textit{Total correct entites in the document}} \quad (1)$$

PRECISION is defined as the number of correct entities retrieved to the total number of incorrect and correct entities retrieved. It is showed as a percentage.

Precision:P=

$$\frac{\textit{Correct entities detected by system}}{\textit{Total incorrect and correct entites detected by system}} \quad (2)$$

$$\text{F- score : F} = {}^{2PR}/_{(P+R)} \quad (3)$$

The dataset is taken from the children story and news article domain. We have taken the children story from (http://abhivyakti-hindi) and the news from (http://webduniya//hindi_news) .The results of F-score are shown in [Table1]:-

The exactness of the system is calculated by the Hindi language experts. From the table 1 it is noticed that pronouns are ambiguous to person, number and gender features While some pronoun can refer to both male and female. These all features affect the performance and F score. . The F score of the news article contains is 86% and The F score is 72 %. The F-score of the overall system based on gender and number agreement is 79%

It is examined that the F-score varies with the structure of sentences. The datasets are complex and narrative style and Hindi is free order. So it affects the combine rules of gender and number agreement .It is also observed that sometimes, demonstrative pronouns (वह , य), Relative pronouns (जसिमें) ,second person pronouns are not resolve correctly and It is observed that certain pronouns refer to both male and female which results the referring to wrong antecedent.

7. Conclusion

Finally to summarize, this paper presents the results of anaphora resolution in Hindi language using rule based method. Table1 shown that the overall system performance in terms of F score is 79% . The proposed system produced better results .Though the system performance is dependent on the structure of the sentences as Hindi language does not have any standard structure. This paper illustrates how gender and number agreement contributes to the performance of anaphora. In this paper it is also presented the experiment that is conducted on the different data set to resolve the anaphora resolution . In this gender and number agreement is taken as a constraint sources which are the base line of our system. The system is formed to determine the contribution of these constraint sources to pronoun resolution on different styles of written text. However, apart from gender and number, coreference resolution, recency, animistic also play important role in anaphora resolution. In the future we will try to include all constraint sources to further increase the performance.

8. References

1. Pal TL, Dutta K, Singh P. Anaphora Resolution in Hindi: Issues and Challenges. International Journal of Computer Applications. 2012 Mar; 42(18).
2. Lakhmani P, Singh S. Anaphora Resolution in Hindi Language. International Journal of Information and Computation Technology. 2013; 3:609–16.
3. Singh S, Lakhmani P, Mathur P, Morwal S. Anaphora Resolution In HINDI Language Using Gazetteer Method. International Journal on Computational Sciences and Applications IJCSA. 2014 Jun; 4:567–9.
4. Chopra D, Purohit GN. Handling ambiguities and unknown words in named entity recognition using anaphora resolution. International Journal on Computational Sciences and Applications IJCSA. 2013 Oct; 3:456–63.

Table1. Result

Data set	No of sentences	Total words	No of pronouns	Resolved pronoun	Correctly resolved pronoun	F-score
News	12	137	7	7	6	86%
Children story	34	445	28	22	18	72%

5. Dutta K, Kaushik S, Prakash N. Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items. The Prague Bulletin of Mathematical Linguistics. 2011 Apr; 95:33–50.
6. Devi SL, V Sundar Ram V, Rao PRK. A Generic Anaphora Resolution Engine for Indian Languages. Proceedings 25th International Conference on Computational Linguistics, Coling. 2014. p. 67–84.
7. Dakwale P, Mujadia V, Sharma DM. A Hybrid Approach for Anaphora Resolution in Hindi Praveen. Proceedings 6th International Joint Conference on Natural Language Processing, IJCNLP, Nagoya, Japan. 2013 Oct 14-18. p. 80–6.
8. Uppalapu B, Sharma DM. Pronoun Resolution For Hindi. Proceedings DAARC2009. 2009 Apr 22; 5847.
9. Sinha RM, Jain A. AnglaHindi: an English to Hindi machine-aided translation system. MT Summit IX, New Orleans, USA. 2003 Sep 23:494–7.
10. Dutta K, Prakash N, Kaushik S. Resolving pronominal anaphora in hindi using hobbs algorithm. Web Journal of Formal Computation and Cognitive Linguistics. 2008 Jan; 1(10):5607–11.
11. Kaur S, Mohana R. A roadmap of sentiment analysis and its research directions. International Journal of Knowledge and Learning. 2015; 10(3):296–323.
12. News Article. Available from: http://webduniya//hindi_news Children Story. Available from: http://abhivyakti-hindi