International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 7 (2013), pp. 609-616

http://www. irphouse.com /ijict.htm

Anaphora Resolution in Hindi Language

Priya Lakhmani[1] and Smita Singh[2]

[1]Department of Computer Science, Banasthali University C-62, Sarojini Marg, C-Scheme, Jaipur, India.

Abstract

In this paper we present our report on anaphora resolution for Hindi language. Anaphora resolution is a key problem in natural language processing, and has correspondingly received a significant amount of attention in the literature. The primary focus of this work is the resolution of pronominal anaphora means binding of pronoun with their intended noun phrase in the discourse. Though the significant amount of work has been done in English and other European languages, the efficient work, in Hindi language, is lagging far behind. The complete paper is divided into four sections. First section of the paper presents a review of work done in the field of anaphora resolution in Hindi language. In the next section we cover issues related to syntactic and semantic structure of Hindi and influence of cases on pronouns. Further we define constraint sources which will form the base of anaphora resolution task. Finally we perform manual experiment on different kinds of data sets and corresponding results are obtained which shows final accuracy of approx 71%.

Keywords: Anaphora, Pronominal Resolution, Case Marker, Natural Language Processing.

1. Introduction

Pronominal or anaphora resolution is defined as the problem of determining the noun phrase (NP) that refers to a pronoun in a document. The 'pointing back' word or phrase is called anaphor. The entity to which an anaphor refers or for which it stands is its antecedent. So in simple term anaphora resolution is the process of determining the antecedent of anaphora. In the following sentence,

"Ram ne shyam ko uski pustak di".

Here ‘uski’ is a pronoun which refers to the noun ‘Shyam’.

A human can quickly work out that in above example, the pronoun ‘uski’ refers to ‘Shyam’. The underlying process of how this is done is yet unclear, especially when we encounter more complex sentences:

S1: “bacchon ne kele khaye kyunki ve bhukhe the”.
S2: “bacchon ne kele khaye kyunki ve pake hue the”.

In sentence S1‘ve’ refers to ‘bacchon’ whereas in sentence S2 ‘ve’ refers to ‘kele’. This is an example of pronominal resolution. An important problem in natural language processing is the resolution of pronouns to their intended referents. This is a difficult task to be handled by an anaphora resolution system. Consequently, anaphora resolution presents a challenge, and is an active area of research. The most common type of anaphora is the pronominal anaphora and the major classifications in pronominal are the first, second and third person pronouns.

Classification of anaphora and pronoun in Hindi language:

Hindi language is a free word order. Pronoun in hindi exhibits a great deal of ambiguity. Pronoun in the first, second, and third person do not convey any information about gender. In Hindi there is no difference between ‘he’ and ‘she’. ‘veh’ is used for both the gender and is decided by the verb form. With respect to number marking, while some forms, like ‘usko’(him), ‘usne’(he) are unambiguously singular but some forms can be both singular and plural, like ‘unhone’ (he)(honorific)/they, or ‘unko’(him)(honorific)/ them. The summary of comparison of pronominal anaphora for third person paradigm in English and Hindi:

Table 1: Pronominal features in English and Hindi for third person paradigm.

Pronominal anaphora in English	**Pronomial anaphora in Hindi**
He	veh
He, She (honorific)	Ve inhon-ne
His, her, its	us
Him, her	usko
This	yeh
That	veh
They These	ve
Them	unko / unse
Their	unka/unki/unke

Himself Herself Itself Themselves	apne swayam khud -aap apne/apni/apnee

2. Related Work

In Hindi and other Indian languages anaphora resolution studied are presented by Bharti et al. [1]. Authors designed methods to handle anaphora and implemented in a prototype natural language interface (NLI) for Hindi. Parse structure of sentence has been formed by Panini parser developed at IIT Kanpur. Sobha and Patnaik[2] gave a rule based approach for the resolution of anaphora in Hindi and Malayalam. Anaphora resolution using rule based approach, corpus based studies, and using centering theory are presented in [3]-[4]. Prasad's thesis work is based on the principle that the grammatical function is important for discourse salience in Hindi Language, as in [3]. Dutta et al. [5] presented modified Hobbs algorithm for Hindi. The algorithm takes into account the free word-order and grammatical role in pronoun resolution in Hindi. Authors concluded that in Hindi, the role of subject and object are significant for reflexive and possessive pronouns. Dutta et al. [6] also highlighted the importance of anaphora resolution for machine translation application by evaluating the existing Machine translation systems: AnglaHindi by IIT Kanpur, Matra2 by CDAC Mumbai and Google translation system. The work of [7], [8] studied the application of machine learning algorithms and probabilistic neural network models on the demonstrative pronouns in Hindi. The work conducted so far, in [7], [8]; demonstrate that classification of demonstrative pronouns as direct and indirect anaphora is essential for successful anaphora resolution. The work is conducted on the Emille Corpus. The studies conducted so far, as in [6], demonstrate that, for a successful NLP application the resolution of anaphora is essential.

3. Issues and Challenges

Resolving anaphora in hindi is a complex task. There are certain issues which are needed to be considered while performing anaphora resolution. These are mentioned below:

- *Encoding in standard form:* Large amount of information is available in Hindi on www (on electronic document form). But this information is encoded in different fonts. That is, there is difficulty in encoding the document in some standard form. Unicode might be a solution to this problem of standardization.
- *Requirement of Unicode based tools for Hindi:* The problem with Unicode based font is that Unicode based tools may not support Hindi. This lack of standardization limits the use of these documents in developing corpus. Therefore, neither a single corpus nor a language processing tool is developed

and freely available for research. The tools available are either not up to the mark or limited to some specific domain only.

- *Pleonastic 'it':* Translation of pleonastic 'it' from English to Hindi creates big difficulty. For example, consider the sentence
- "It is raining heavily today"
- It has corresponding translation in Hindi as "*aaj tei baarish ho rhi hai*". Though the corresponding translation of 'it' in Hindi be '*yeh*' or '*veh*', in the given example it have no mapping. Therefore it is quite irrelevant to translate this type of "it" in Hindi target text form English source text. Frequent occurrences of this type of 'it' can cause problem in machine translation[6].
- *Cases and their influence:* Hindi does not differentiate pronouns on gender, its verb that differentiate masculine from feminine gender. Therefore knowledge of verb is also essential for correct pronoun resolution. In Hindi, cases plays very important role in correct translation of some source text in some foreign language to target text in Hindi. The case marker is added separately and the pronoun modifies accordingly. The agreement inflection is marked for person, number, gender.

Table 2: Role of verb phrase in gender disambiguation for pronoun 'veh' (he/she).

English Sentence	**Hindi Sentence**	**Observation**
He is happy	veh khush hai	No Gender differentiation
She is happy	veh khush hai	No Gender differentiation
He was happy	veh khush **tha**	Gender is obtained from tha (male) , thii(female)
She was happy	veh khush **thii**	Gender is obtained from tha (male) , thii(female)

4. Constraint Resources

An experiment based on anaphora resolution has been conducted by us for which these constraint sources forms the base line. Originally only semantic constraint sources were going to be used. However syntactic constraint sources are included because they include some of the most effective techniques relative to their difficulty to implement. The modules available for use are:

- *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. This is an example of a constraint that relies on accurate part of speech tagging in the preprocessor.
- *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.

5. Experiment and Result

We have performed two experiments on different types of data sets. This first experiment used text from a children's story. Ideally this experiment represents a baseline performance since the story is a straightforward narrative style with extremely low sentence structure complexity. 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 perform anaphora approach manually over these stories. Another experiment is conducted on news articles from IBN khabar in Hindi language.(http://khabar.ibnlive.in.com/tag/tag_topic/all/Dowry-Case.html). It presents an entirely different challenge from the narrative story style. From the experiment following accuracy was observed:

Table 3: Result from experiment 1 and 2 performed on short stories.

	Experiment 1			**Experiment 2**		
Constraints	Correctly resolved	Anaphora to resolve	Accuracy	Correctly resolved	Anaphora	Accuracy
1.Recency	33	77	42.87%	26	52	50.00%
2.Recency, Number agreement	37	77	48.05%	28	52	53.80%
3.Recency, Number, Gender agreement	37	77	48.05%	29	52	55.76%

4.Recency, Number agreement, Gender agreement, Animistic knowledge	55	77	71.44%	37	52	71.10%

The result shows that Number Agreement increased accuracy to 5.18% and Gender Agreement had no contribution to accuracy whereas animistic knowledge increases the accuracy by 23.39% in Experiment 1.

The result of Experiment 2 shows that recency provides 50% accuracy which proves that recency is a baseline criteria for anaphora resolution in hindi language. Next, the number agreement and gender agreement shows a little improvement in accuracy. Further, animistic knowledge contributes significantly to overall accuracy thereby increasing it to 71%.

6. Conclusion

This paper presents the brief description of anaphora resolution in Hindi language. Hindi language is free word order and hence it has several complications in resolving pronoun in compare to English language. A manual experiment resolving anaphora is performed manually on different data sets. Several constraints are considered which forms the base line of our experiment. The experiment is conducted to determine the contribution of different constraint sources to pronoun resolution on different styles of written text. In future we will try to pair more constraint sources with the writing styles for which they contribute the most to the accuracy of the pronoun resolution system.

References

[1] A Bharati, Y Krishna Bhargava and R. Sangal (1993), "reference and ellipsis in an Indian languages interface to database," *computer science and i**nformatics*, IIT Hyderabad, **23**; 3, pp.60.

[2] L. Sobha and B.N. Patnaik (2002), "Vasisth: An anaphora resolution system for Malayalam and Hindi", *Symposium on Translation Support Systems,*

[3] R. Prasad (2003), "Constraints on the generation of referring expressions: with special reference to Hindi," (Ph.D. thesis), University of Pennsylvania,

[4] R. Prasad and M (2000)., "Discourse salience and pronoun resolution in Hindi," *In Williams, A. & Kaiser, E. (eds.) Penn* Working Papers in: Current Linguistics Work in Linguistics, **6**, 3, pp.189-208.

[5] K. Dutta, N. Prakash and S. Kaushik (2008), "Resolving Pronominal Anaphora in Hindi using Hobbs‟ algorithm," *Web Journal of Formal Computation and Cognitive Linguistics*, **10.**

[6] K. Dutta, N. Prakash and S. Kaushik (2009), "Application of Pronominal Divergence and Anaphora Resolution in English-Hindi Machine Translation," *Research journal "POLIBITS" Computer Science and Computer Engineering with Applications*, **39**, pp-55-58.

[7] K. Dutta, N. Prakash and S. Kaushik (2010), "Probabilistic Neural Network Approach to the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi," *Expert Systems with Applications*: An International Journal, **37**, 8, pp. 5607-5613,

[8] K. Dutta, S. Kaushik and N. Prakash (2011) , "Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items," *Prague Bulletin of Mathematical Linguistics*. Versita, **95**, pp. 33-50.