Event Anaphora Resolution in Natural Language Processing for Hindi text

Komal Mehla¹, Karambir¹ and Ajay Jangra¹

¹ Computer Science & Engineering Department, University Institute of Engineering and Technology, Kurukshetra, Haryana, India

Abstract
This paper presents a comprehensive study about the anaphora resolution. Anaphora resolution can be of any type such as Entity Anaphora resolution and Event Anaphora resolution etc. Event Anaphora resolution is the main focus of this review paper. This resolution of Event anaphora can be different for different languages as the structure of the language changes. Here the literature about anaphora resolution and the problems that encountered in resolution of Hindi text are discussed.

Keywords: (Natural Language Processing), AR(Anaphora Resolution).

1. Introduction
In order to support new educational paradigms, decision making and business competitiveness in the age of electronic mail and internet, there is a need to process and assimilate vast amount of on-line information. Computational linguistics is one of the many growing fields in recent years. The application areas related to it like text summarization, information extraction and information retrieval have attracted a lot of research. Anaphora Resolution is also known as Co-reference resolution. It is a branch of Information Extraction, which has attracted a much attention. The core aspect of semantic processing is Anaphora resolution. Simply it is the process of automatic identification of anaphors, for example pronouns and their antecedents (the entities to which anaphors refer). Anaphora resolution can be of these types.

1.1 Entity Anaphora Resolution
Entity anaphora stands for those pronominal references which refer to a Concrete Entity such as Person, place and other common nouns. Thus possible candidate referents for entity anaphora are noun phrases (NP).

1.2 Event Anaphora Resolution
Event anaphora stands for those pronominal references which refer to Events Thus possible candidate referents are verbs, clauses and propositions. Anaphora Resolution in Hindi is a complex task. While performing anaphora resolution certain issues should be considered.

2. Literature Review
In the field of anaphora resolution most of the work is done in English and other European Languages. Hirst et al. [1], proposed a computational solution for the co-reference and anaphora resolution in English. An effective algorithm for anaphora resolution is Hobb’s algorithm, as in[2][3][4]. Instead of semantic information it uses syntactic information. The Hobb’s algorithm[4] depends on number checker, morphological gender and syntactic parser. Due to this, when evaluating new pronominal anaphora resolution algorithm, it is often used as a baseline. Mitkov’s et al. [5] and JavaRAp [6] worked well towards anaphora resolution and system proposed by them are considered under good anaphora resolution systems. Earlier manually developed knowledge by linguistics were used in knowledge based anaphora resolution approach. Hearer’s set of beliefs(knowledge) solve this type of anaphora resolution, which is used to represent discourse algorithm, semantic and syntax. Manually processed inputs are required by these approaches. Normally it is assumed that inputs are perfect and are checked or modified by the experts, if required. Pre-processing tools’ existence and the difficulties arrived in adapting the knowledge based system for other languages motivate the researchers to focus into knowledge poor approaches (heuristics based systems). Robustness and simplicity are provided by Knowledge Poor (KP) systems. In performance these systems are comparable to Knowledge Based systems.
The various approaches to resolve anaphora are:

3. Approaches to anaphora resolution

The various approaches to resolve anaphora are:
3.1 Centering Theory

Centering theory is a discourse based approach. To model what a sentence is speaking about a framework is given by this theory. The framework can be used to identify that the pronouns are referring to which entities in a given sentence. Attentional salience of discourse entities are modeled in this theory. From the computational point of view centering based approaches are more attractive. It is because they obtained the required information from the properties of utterances alone.

3.2 Lappin Leass algo

Lappin Leass algo is a hybrid method which uses a model. On the basis of different factors the discourse salience of the candidate is calculated by the model. Different weights are given to the factors according to their relevancy. While evaluating antecedents cost semantic or real world knowledge were not used by them.

3.3 Gazetteer method

External knowledge is provided to the system by Gazetter method, by creating lists. Based on certain operations elements of lists are classified. Therefore it is also known as List Lookup Method.

4. Issues and Challenges

Issues and challenges for anaphora resolution are mentioned below:

4.1 Encoding in standard form

On www (on electronic document form) large amount of information is available in Hindi. But different fonts are used to encode this information. Hence, there is difficulty in encoding the document in some standard form. This problem of standardization might be solved by unicode.

4.2 Requirement of Unicode based tools for Hindi

Unicode based tools may not support Hindi and it is the problem with unicode based font. Use of these documents in developing corpus is limited due to this lack of standardization. Therefore, neither a language processing tool nor a single corpus is developed. Priya Lakhmani and Smita Singh 612 are freely available for research. The tools available are either limited to some specific domain only or not up to the mark.

4.3 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 tej baarish ho rhi hai”.

The corresponding translation of ‘it’ in Hindi be ‘yeh’ or ‘veh’, but in the given example it have no mapping. Hence it is irrelevant to translate this type of “it” in Hindi target text form English source text. Problem in the machine translation can be caused due to frequent occurrences of this type of ‘it’.

4.4 Cases and their influence

Hindi does not differentiate pronouns on gender, its verb that differentiate masculine from feminine gender. For correct pronoun resolution the knowledge of verb is also essential. In correct translation of some source text in some foreign language to target text in Hindi cases plays very important role. The case marker is added separately and the pronoun modifies accordingly. For person, number, gender the agreement inflection is marked.

5. Conclusion

As discussed earlier, there are many algorithms which were used for anaphora resolution. These algorithms are different according to the language they support. Some algorithms failed in the main factor to determine the performance of an algorithm, which is modularity. A new machine learning based approach for event anaphora resolution that will try to overcome these problems can be developed.

References


