IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 1, January 2015.
www.ijiset.com
ISSN 2348 – 7968

Event Anaphora Resolution in Natural Language Processing for Hindi text

Komal Mehla[1], Karambir[1] and Ajay Jangra[1]

[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 ayntax. 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.

ISSN 2348 – 7968

Models/ Element	***MARS***	***RAP***	***MOA***	***Jepthah***	***ARN***
Data Sets	Technical Manual	Computer Manual	Technical Manual	Manual Dialogues, G. B. Shaw Play	Oslo and BREDT corpus
Prerequisite	Pre-processing task Conducted automatically using FDG parser	The data manually check and corrected Pleonastic it manually remove	The data manually check and corrected Pleonastic it manually remove	manually tagged corpora, then be tuned using Genetic Algorithms	manually tagged corpora, then be tuned using Genetic Algorithms
Purpose	Third personal pronouns and lexical anaphora (reflexives and reciprocals)	Third personal pronouns and lexical anaphora (reflexives and reciprocals)	Third personal pronouns and lexical anaphora (reflexives and reciprocals)	pronouns, reflexives and deictic anaphora	Third personal Pronoun exclude 'it'
Language	English	English	English	English	Norwegian
Reported success rate	61.55	84.10	89.70	72.50	70.50

Table 1: Comparison of existing systems

Mitkov's Original Approach (MOA) [5], RAP[6], ARN[7] are some of the examples of Knowledge Poor approach. [8][9] gives brief comparison of these systems. The approach used by Agarwal et al. [10] is based on matching constraints for the grammatical attributes of different words. The important point about this approach is that along with their linguistic information.

They use animate/non-animate classification of entities while resolving the reference.

Though the approach claims accuracy 96% for simple sentences and 80% for complex sentences, it neither gives any detail about the Coverage of individual constraints or features, nor any detail about the data except that it is taken from children stories. Also it claims results on only 120 pronouns, hence reproducing or verifying the results and hence establishing the validity of the algorithm is difficult.

Dutta et al. [11] present an approach in which they propose adopting Hobb's algorithm for Hindi as a baseline algorithm. However, the algorithm is implemented over a few sentences for a limited types of pronoun and no results are reported. They state that the algorithm can be further evaluated and used once a sufficient amount of data is available, however, phrase structure data for Hindi in sufficiently large amount is yet unavailable.

The most important related work for anaphora resolution in Hindi is presented by Prasad and Strube, [12]. Discourse salience ranking to two pronoun resolution algorithms, the BFP and the S-List algorithm is applied by this algorithm. This approach is mostly based on Centering theory.

Ritual Iida et al. [13] developed a new ILP-based model of zero anaphora detection and resolution that extends the co-reference resolution model proposed by Denis and Baldridge by introducing modified constraints and a subject detection model.

C.A. Bevan et al. [14] described a new class of non-parametric, unsupervised Bayesian models which were designed for the purpose to solve the problem of event co-reference resolution. Specifically, they have showed that in order to relax some of the limitations how already existing models could be extended and how to represent the event mentions from a particular document collection better. In this regard, models can be devised for which the data could be used to infer the number of clusters and the number of feature values corresponding to event mentions automatically.

Zhang et al.[15] presented a framework based on machine learning for event pronoun resolution which explore both flat and structural features. Random Down Sampling Method [16] was used to overcome the problem of class imbalance between negative and positive classes. Testing instances could also be generated in the similar way. Since the negative instances are far more than the positive instances, so the verb candidates are chosen as- the main verb of the previous two sentences and the main verbs of the clauses after them, in the current sentence the main verb (if it is before anaphor) and the main verb of the clauses before it were selected.

Donna K. Byron[17] gave a technique named PHORA to resolve pronominal reference to individual or abstract entities. It applies semantic filtering in addition to salience calculations, due to which it was able to resolve pronouns referring to less salient abstract entities such as actions, prepositions and kinds. PHORA can work equally on personal
(it, him) and demonstrative pronouns (this, those etc.). 72% of test pronouns are successfully interpreted by it as compared to 37% for a leading technique not having these features.

3. Approaches to anaphora resolution

The various approaches to resolve anaphora are:

ISSN 2348 – 7968

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

[1] G. Hirst, “Anaphora in Natural Language Understanding,” Springer-Verlag, Berlin, 1981.

[2] D. Jurafsky and J. Martin, “Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition,” Upper Saddle River, NJ: Prentice Hall, 2010.

[3] R. Mitkov, "An integrated model for anaphora resolution", 15th conference on Computational linguistics - Volume 2 Kyoto, Japan: Association for Computational Linguistics, 1994.

[4] J. R. Hobbs,J. R., “ Pronoun Resolution” ,Research Report 76-1, Department of Computer Science, City College, New York., August 1976.

[5] R. Mitkov, R. Evans and C. Orasan, “A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method,” LECTURE NOTES IN COMPUTER SCIENCE, 2002, ISSU 2276, pp.168-186.

ISSN 2348 – 7968

[6] S. Lappin and H. Leass, "An algorithm for pronominal anaphora resolution," Computational Linguistics, 20(4), 535-561,

[7] G. I. Holen, "Automatic Anaphora Resolution for Norwegian (ARN)," in Department of Linguistics and Scandinavian Studies. vol. Master Norway: UNIVERSITY OF OSLO, 2006, p. 142.

[8] N. K. M. Noor, M. J. A. Aziz, S. A. Noah and M. P. Hamzah, "Anaphora resolution of Malay Text: Issues and Proposed Solution Model," International conference on Asian Language Processing, 2010.

[9] P. Jain, M. Mittal, A. Mukerjee and A. Raina, "Anaphora Resolution in Multi-Person Dialogue, Strube, M. and Candy Sidner (ed.)," Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Association for Computational Linguistics, Boston, Cambridge, Massachusetts, 2004.

[10] S. Agarwal, M. Srivastava, P. Agarwal, and R. Sanyal. "Anaphora resolution in hindi documents" In Natural Language Processing and Knowledge Engineering, 2007(NLP-KE 2007) International Conference on, pages 452–458. IEEE, 2007.

[11] K. Dutta, N. Prakash, and S. Kaushik. "Resolving pronominal anaphora in hindi using hobbs algorithm" Web Journal of Formal Computation and Cognitive Linguistics, 1(10), 2008.

[12] R. Prasad and M. Strube. "Discourse salience and pronoun resolution in hindi" U. Penn Working Papers in Linguistics, 6:189–208, 2008.

[13] R. Iida and M. Poesio "A Cross-Lingual ILP Solution to Zero Anaphora Resolution" Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 804–813, Portland, Oregon, June 19-24, 2011.

[14] C. A. Bejan and S. Harabagiu "Unsupervised Event Coreference Resolution" in Association for Computational Linguistics , volume 40, issue 2, 2014.

[15] Z. Ning, K. Fang and L. Peifeng "Research of Event Pronoun Resolution" in International Conference on Asian Language Processing, 2011.

[16] M. Kubat and S. Matwin "Addressing the curse of imbalanced data set: one sided sampling" In proceeding of 14th International Conference on machine learning, 1997.

[17] K. Donna, Byron " Resolving Pronominal Reference to Abstarct Entities" Proceeding of 40th annual meeting of the Association for Computational Linguistics , july 2002.