Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.576 – 580

RESEARCH ARTICLE

Performance Analysis of two Anaphora Resolution System for Hindi Language

Priya Lakhmani[1], Smita Singh[2], Sudha Morwal[3]
Department of Computer Science, Banasthali University, India
[1] tinalakhmani@gmail.com; [2] smitasingh101@gmail.com; [3] sudha.morwal@yahoo.com

Abstract— *One of the challenges in NLP is to determine what entities are referred to in the discourse and how they relate to each other. This is known as Anaphora resolution. Basically there are three main algorithms for anaphora resolution- Hobbs, Centering and Lappin Leass algorithm. This paper presents the comparison of two computational models for resolving anaphora. The first model is based on the concept of Lappin and Leass algorithm and the second model is based on the concept of Centering algorithm. Both of these model works on Hindi language. As Hindi language is quiet complicated with respect to other European languages, there are many factors needed to be considered for resolving anaphora. Our computational models uses Recency factor as a salient factor. An experiment is conducted on short Hindi stories and the comparative result for both the models is summarized. The respective accuracy for both the model is analyzed and finally the conclusion is drawn for the best suitable model for Hindi Language.*

Keywords: Anaphora; Centering algorithm; Discourse; Lappin Leass algorithm; Natural language processing

I. INTRODUCTION

In linguistics, the term anaphora denotes the act of referring. It is the use of an expression, the interpretation of which depends upon another expression in context. Reference to an entity that has been previously introduced into the discourse is called anaphora. Any time a given expression refers to another contextual entity, anaphora is present. The entity referred back to is called the 'referent' or 'antecedent'. Anaphora denotes the act of referring to the left, that is, the anaphor points to its left toward an antecedent (in languages that are written from left to right). The process of binding the referring expression to the correct antecedent, in the given discourse, is called anaphora resolution. Pronoun resolution involves binding of pronouns to the correct noun phrase. Consider the sentence:

"सीता मेले में घूमने गयी | उसने वहाँ फल खाये |"

This sentence demonstrates an anaphora, where the pronoun 'उसने' refers back to a referent. Intuitively, 'उसने' refers to 'सीता'. The pronoun' वहा' refers to 'मेले.'

Anaphora resolution can be intrasentential (where the antecedent is in the same sentence as the anaphor) as well as intersentential (where the antecedents are in a different sentence to the anaphor). When performing anaphora resolution all noun phrases are typically treated as potential candidates for antecedents. The scope is usually limited to the current and preceding sentences and all candidate antecedents within that scope are considered.

A. Classification of anaphora and pronoun in Hindi:

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'. 'वह' is used for both the gender and is decided by the verb form. With respect to number marking, while some forms, like 'उसको'(him), 'उसने'(he) are unambiguously singular but some forms can be both singular and plural, like 'उन्होने' (he)(honorific)/they, or 'उनको'(him)(honorific)/them. Hence Resolving anaphora in Hindi is a complex task.

II. RELATED WORK

An extensive research work has been done in English and European language. Related work for anaphora resolution is broadly classified by three main algorithm developed by researchers.

- First work in the field of pronoun resolution is done by J.R Hobbs in English language in 1976. Hobb's algorithm makes use of syntactic information for resolving pronoun. It gave accuracy of 82% for English language.
- Joshi, A. K. & Kuhn. S, in 1979 and Joshi, A. K. & Weinstein.S in 1981, gave centering theory for pronoun resolution. This work is also done in English language which gave 76% accuracy[10].
- S. Lappin and H. Leass proposed their algorithm for pronoun resolution for English language in year 1994. Lappin and Leass evaluated RAP (Resolution for anaphora procedure) using 360 pronoun finds the correct antecedent for 310 pronouns, 86% of the total (74% of intersentential cases and 89% of intrasentential cases)[1].

Futher, the work is done by researchers based on the concept of above mention algorithms. The work done for anaphora resolution based on Lappin Leass algorithm is summarized below:

- Using Lappin Leass approach pronominal anaphora is resolved in Nepali Language by Dev Bahadur[5].
- Work also has been done in Portugeese language using Lappin and Leass algorithm by Thiago Thomes Coelho, Ariadne Maria Brito Rizzoni Carvalho[1].
- Anaphora resolution in Spanish text is done by M Palomar using Lappin Leass[6].
- In 1990, S.Lappin and M.McCord developed a syntactic filter on pronominal anaphora for slot grammer using Lappin Leass principles[4].
- Christopher Kennedy and Branimir Boguraev used the concept of Lappin Leass algo and resolve anaphora without using a Parser[2].
- Anaphora is resolved in Multi-Person Dialogues using Lappin Leass algo by Prateek Jain , Manav Ratan Mital , Sumit Kumar , Amitabha Mukerjee and Achla M. Raina[3].

The work done for anaphora resolution based on Centering algorithm is summarized below:

- Manuel Palomar, Lidia Moreno and Jesfis Peral resolved anaphora in Spanish Texts using Centering approach[6].
- In 1998, Hoffman &Beryl present a Word order, information structure using Centering principles in Turkish language[8].
- In 1999, Strube & Hahn present a system for anaphora resolution for German based on extension of Centering theory[9].

- C .Navarretta resolved anaphors in Danish text by using Centering theory.
- Work also has been done in Portugese for resolving pronouns by Fernando José Vieira da Silva, Ariadne Maria Brito Rizzoni Carvalho, Norton Trevisan Roman using Centering principles[11]
- Centering principles also applied in Tamil by J.Balaji[7].

III. RESOLVING SYSTEM

Based on the literature survey it is analyzed that a huge amount of work has been done in European languages. Both Lappin Leass algorithm and Centering algorithm aren't implemented for Hindi Language. So we developed two computational models for resolving anaphora in Hindi Language. The first model is based on the concept of Lappin and Leass algorithm.

A. Lappin and Leass algorithm:

It falls under the category of hybrid approach. It use a model that calculates the discourse salience of a candidate based on different factors calculated dynamically and use this salience measure to rank potential candidates. The factors that the algorithm uses to calculate salience are given different weights according to how relevant the factor is.

B. Centering algorithm:

The second model is based on concept of Centering algorithm. Centering theory provides a framework to model what a sentence is speaking about. This idea can be used to find which entities are referred to by pronouns in a given sentence. This theory models the attentional salience of discourse entities, and relates it to referential continuity. Centering has certain transitions rule based on which it resolves anaphora. Centering uses salience factors internally on these transition rules.

As Hindi language has a different grammar from that of European language, there are certain factors needed to be considered for correct pronoun resolution.

C. Factors responsible for resolving pronoun in Hindi language:

- *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

The computational model based on the above approaches uses Recency factor for resolving anaphora. Recency factor describes that the referents mentioned in current sentence tends to have higher weights than those in previous sentence. For example consider the sentence,

"अमित ने कृष्णा को आम दिया | वह कच्चा था |"

In this sentence there are three nouns 'अमित', 'कृष्णा', 'आम'. According to Recency factor the highest weight is assigned to the most closest noun 'आम' from the pronoun 'वह'.

Based on Recency factor the experiment is conducted on short documents to measure the overall system performance. The Recency factor in both the models gives approximate 40 to 50 percent accuracy. Other factors may be added to increase the accuracy.

D. Difference in two computational models:

As both the model uses Recency factor, still there is a large difference in the techniques for resolving pronoun.

- Lappin and Leass algorithm uses mathematical calculation for pronoun resolution whereas Centering algorithm uses transitions without using any calculation.
- Lappin and Leass algorithm is a part of Hybrid approach for anaphora resolution whereas Centering algorithm is a Discourse based approach.

IV. EXPERIMENT AND RESULT

A standard experiment is based on finding the contribution of Recency factor to the overall accuracy of correctly resolved pronouns. We have performed experiment on the same data sets by two different computational models. In both of these models Recency is a salience factor. Based on Recency, accuracy is calculated. The difference in accuracy of both model is analyzed.

A. Data Set:

This experiment used the text from a children's story. 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 anaphora resolution over these stories. Ideally this experiment represents a baseline performance since the story is a straightforward narrative style with extremely low sentence structure complexity. Also it contains approx 10 to 25 sentences having 100 to 300 words.

The first computational model based on Lappin and Leass algorithm shows the following result:

Table I. Result of model form concept of Lappin and Leass algorithm

Data Set	**Total Sentences**	**Total Word**	**Total Anaphors**	**Correctly Resolved Anaphor**	**Accuracy**
Story1	12	133	11	5	45%
Story2	11	132	8	5	62%
Story3	23	275	20	4	20%
Story4	17	213	17	10	58%
Story5	21	227	20	10	50%

The result of this experiment shows that Recency provides approx 50% accuracy which proves that Recency is a baseline criterion for anaphora resolution in Hindi language.

The second computational model based on Centering algorithm shows following results:

Table II. Result of Model based on concept of Centering algorithm

Data Set	Total Sentences	Total Word	Total Anaphors	Correctly Resolved Anaphor	Accuracy
Story1	12	133	11	4	36%
Story2	11	132	8	3	37%
Story3	23	275	20	3	15%
Story4	17	213	17	8	47%
Story5	21	227	20	7	35%

The result of Centering algorithm shows that Recency factor provides approx 35% accuracy to overall system.The correctness of the accuracy obtained by the experiment is measured by the language expert. From the above results it is concluded that for Hindi language Lappin Leass algorithm is more suitable as compared to Centering algorithm as it gives better accuracy. Also in Centering algorithm the transition rules depends on word ordering of sentence, hence it affects the accuracy. It is also observed that both the system fails to differentiate between animate and Inanimate things. Further Recency factor proves to give approx 50% accuracy for anaphora resolution which signifies that Recency is the base factor. In order to increase accuracy more factors can be added such as gender agreement, number agreement, animistic knowledge etc.

V. CONCLUSION

This paper presents comparative result of two computational models. A standard experiment is performed on same data set in Hindi Language by both the model. The data set initially contains 100 to 300 words. The approximate accuracy of both the model is compared. The experiment is conducted taking Recency factor as a baseline and it gives approximate 50% accuracy. The remaining 50% pronouns are not resolved correctly because the system fails to differentiate between animate and inanimate things. Hence, Recency factor alone is not sufficient for complete correct pronoun resolution system. Other factors also play significant role. So there is a scope for other factors to be added. In future we will try to use animistic knowledge along with Recency factor in order to increase the accuracy of overall system. Furthermore we will try to use a corpus of at least 1000 words to test our computational model.

REFERENCES

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

[2]Kennedy, Branimir Boguraev, “Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser”, University of California

[3]Prateek Jain and Manav Ratan Mital and Sumit Kumar and Amitabha Mukerjee and Achla M. Raina “Anaphora Resolution in Multi-Person Dialogues”, Indian Institute of Technology Kanpur.

[4]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.

[5]Dev Bahadur Poudel and Bivod Aale Magar “Anaphoric Resolution in Nepali”, Nepal Engineering College.

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

[7]“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

[8]Hoffman, Beryl. (1998). “Word order, information structure, and centering in Turkish”.

[9]Strube & Hahn (1999) “A system for anaphora resolution for German based on extension of Centering theory”.

[10]Aravind K Joshi, Rashmi Prasad, Eleni Miltsakaki “Anaphora Resolution: A Centering Approach”

[11]Marilyn A Walker, “Centering, Anaphora Resolution, and Discourse Structure”. ATT Labs Research.

