Hybrid Model of Automated Anaphora Resolution

Kalyani Kamune, Avinash Agrawal
Department Of Computer Science, RKNEC, Nagpur, India

Article Info

Article history:

Received May 21, 2014
Revised Aug 7, 2014
Accepted Aug 19, 2014

Keyword:

Anaphora
Anaphora resolution
Antecedent
Discourse Modelling
Pronominal Resolution

ABSTRACT

Anaphora resolution has proven to be a very difficult problem of natural language processing, and it is useful in discourse analysis, language understanding and processing, information exaction, machine translation and many more. This paper represents a system that instead of using a monolithic architecture for resolving anaphora, use the hybrid model which combines the constraint-based and preferences-based architectures, each uses a different source of knowledge, and proves effective on theoretical and computational basis. An algorithm identifies both inter-sentential and intra-sentential antecedents of “Third person pronoun anaphors”, “Pleonastic it”, and “Lexical noun phrase anaphora”. The algorithm use Charniak parser (parser05Aug16) as an associated tool, and it relays on the output generated by it. Salience measures derived from parse tree, in order to find out accurate antecedents from the list of potential antecedents. We have tested the system extensively on 'Reuters Newspaper corpus'.

Corresponding Author:

Kalyani Kamune,
Departement of Computer Science,
Shri Ramdeo Baba College o Engineering and Managment,
Katol Rd, Nagpur, Maharashtra, India.
Email: kalyanikamune24@gmail.com

1. INTRODUCTION

Interpreting anaphoric expressions is one of the most fundamental aspects of language interpretation. The study of anaphora and anaphora has brought about many fundamental developments in theoretical linguistics and computational linguistics and has important practical applications in work on information extraction, summarization, and entity disambiguation. Anaphora resolution is a complicated problem in Natural Language Processing and has attracted the attention of many researchers. Anaphora describes the language phenomenon of referring to a previously mentioned entity (also called object or event); anaphora resolution is the process of finding that previous item. Consider the following clarifying example from a British World War II anti-raid leaflet:

“If an incendiary bomb drops next to you, don’t lose your head. Put *it* in any bucket and cover *it* with some sand.”

If this raised eyebrows - don’t worry - it is meant to. Indeed “it” could stand for (or *refer* to) either of the two objects mentioned before it, “Bomb” and “head”. The authors meant the former, but the rules of language have a tendency to bias readers to picking the latter. But then “head’s” are not the usual things one puts in buckets and covers with sand. What anaphora resolution, when done correctly, enables us and systems to do, is to merge the previous information about an entity with the new information we encounter. So think of anaphora as the intricate balance between conciseness of communication and the ability of humans to understand each other [1].

The remaining sections of this paper are organized as follows. In section 2, we represent the Literature Survey of Anaphora Resolution. In section 3, we represent in details how each step of “Automated Anaphora Resolution” algorithm is implemented. Section 4, explains the architecture of “Automated Anaphora Resolution”.

2. LITERATURE SURVEY

In the process of anaphora resolution, antecedents can be noun or verb phrases, any clauses, any sentence or even paragraphs/discourse segments as antecedents. Finding antecedents as a noun phrase is comparatively an easy task than that of finding rests as an antecedent. Generally, all noun phrases (NPs) preceding an anaphor are initially considered as potential candidates for antecedents. Search limit has to be predefined. An “ideal” anaphora resolution system has the search limit of 17 sentences away from the sentence in which anaphor is present [4].

All the potential antecedents within the search limit preceding anaphora are finding out, and various anaphora resolution factors are used to find the correct antecedent for the particular anaphora. Various factors used can be "eliminatory" i.e. doesn't count certain noun phrases from the set of potential antecedents (such as gender , number , people constraints) or "preferential", assigning more preference to some potential antecedents and less to others (such as salience). [5]

System must define the text segments in which the antecedent can be found, which is called as search limit or anaphoric accessibility space. This step is very important, for further processing. Keeping the search limit too narrow results in the exclusion of valid antecedents and keeping the search limit too broad results in the large candidate lists which are ultimately leads to erroneous results. Once the list of all possible candidates is found, several constraints can be applied in order to remove incompatible antecedents. It may be possible that after applying constraints, some anaphora can still have more than one possible antecedent, in such case we can get accurate antecedent by using preferences.

Two types of architecture are present constraint-based architectures and preferences-based architectures, based on the factors (constraint-based or preferences-based) which we are using in the process of anaphora resolution. Instead of using a monolithic architecture for resolving anaphora, in the “Automated Anaphora Resolution” system we use the combination of constraint-based and preferences-based architectures; each uses a different source of knowledge, and proves effective on theoretical and computational basis. Hence the system mainly works on 3 steps given as: (1) defining an anaphoric accessibility space or Search Limit, (2) apply constraints, and then (3) apply preferences.

3. ‘HYBRID MODEL OF AUTOMATED ANAPHORA RESOLUTION’ IN DETAIL

If we resolve anaphora correctly, it it significantly increases the performance of the downstream Natural Language Processing applications. Hence, to address this problem of resolving anaphora correctly, we have implemented a Java-based system which uses a hybrid approach for resolving anaphora. A system used for identifying both inter-sentential and intra-sentential antecedents of third person pronouns in their nominative, accusative or possessive case and pleonastic anaphora. In our system, we have consider the search limit/ Anaphoric Accessibility Space of 3 sentences, hence for any anaphora, system will find all the potential antecedents from the 3 sentences preceding the sentence in which anaphora is present, including the sentence in which anaphora is present. System uses Charniak parser (parser05Aug16) as an associated tool, and it relays on the output generated by it.

Instead of using a single monolithic architecture system uses the hybrid approach which combines constraint-based and preferences-based architectures System read text file as an input and give it to the sentence splitter. Sentence Splitter used by the system as an associated tool split the sentences and put the tags like <S>and</S> before and after each sentence respectively, according to the requirement of the parser. The output generated by the sentence splitter is then given to the parser05aug16 for further processing. A Syntactic representation is generated by the parser called as parse tree. This syntactic representation created by the parser plays an vital role for the further processing. Next step is to create the list of anaphora and antecedent. Now, each anaphora form a pair with all the potential antecedents comes in its Anaphoric Accessibility Space i.e. Within 3 sentence from the sentence in which anaphora is present. For each pronouns and noun phrase in each pair find agreement features (Number, People and Gender). Each created pair of anaphora and noun phrases is then checks for the agreement feature (Person, Gender, and Number) in agreement filter. If the given pair fulfills the agreement feathers then allow passing for the further processing, else pair is discarded by the system. The resulted pairs are then filtered through further filtering process to get the correct antecedent for the anaphora.

All the information required by the system is not used generated by the parser, system have to extract certain required information from the output generated by the parser. Next step is to derive required salience measures from parse tree, which is used for the further processing. Apply “Pleonastic Pronoun Filter” to find pleonastic pronouns (It will take “List of Anaphors” as an input). Apply ‘Personal Pronoun Filter’ for resolving ‘Third person Pronouns’, which will take list of noun phrases and pronoun as input. Potential Candidates for antecedent are ranked by their “salience weights” and the top one is proposed as the accurate antecedent. Generate output as “Co-referential Pair”.

3.1. Agreement Filter

We get number of pairs of noun and pronoun from the list of noun phrase and pronoun phrase, which are then allowed to filter through ‘Agreement filter’. The agreement features’ compatibility of each and every pair of pronoun and a noun phrase is tested by a *Agreement filter*. It states noun and pronoun pair as non-matching in their agreement features only if at least one agreement feature doesn’t agrees, or else it states them as matching. The value “unknown” is regarded to agree with any value of the feature. The constraint system consists of conditions that must be met, and candidates that do not fulfill these conditions will not be considered possible antecedents for the anaphor.

Consider bellow example:

“Cincinnati Bell Inc said it has started its previously-feather announced 15.75 dlr per share tender offer for all shares of Auxton Computer Enterprises Inc.”

In the above example, 'Cincinnati Bell Inc' and 'it' satisfies the conditions of agreement feature and pass for the father filtering.

1. Required Information Obtained by Inspecting the Parse Tree Structure
   All the information required by the system is not given by the parser; “Automated Anaphora Resolution” system recovers them by using structure information of the verb/noun phrases. [2]
2. Pleonastic Pronouns Filter
   System also identifies pleonastic pronouns. In case of Pleonastic Anaphora, the pronoun doesn't have any referent. System uses the list of modal adjective and a cognitive verb in order to find pleonastic pronouns.
3. Personal Pronoun Filter and Lexical Filter
   “Automated Anaphora Resolution” system is used to find out the “Third person pronouns” in their nominative, accusative or possessive case and “Lexical anaphora”. Information Obtained by Inspecting the Parse Tree Structure plays a vital role for identifying the “Third person pronoun anaphora” and “Lexical Anaphora”. It is the most wide spread type of anaphora. Pronominal anaphora occurs at the level of personal pronouns, possessive pronouns, reflexive pronouns. The set of anaphoric pronouns consist of all third person personal (he, him, she, her, it, they, them), possessive (his, her, hers, its, their, theirs), reflexive (himself, herself, itself, themselves) pronouns in both singular and plural. Whereas, first and second person pronoun can often be non-anaphoric.

System is used to find out the “Third person pronouns” in their nominative, accusative or possessive case. Information obtained from the Parse Tree Structure plays a vital role for identifying the “Third person pronoun anaphora”.

For Example:

“Sumitomo President Koh Komatsu told Reuters he is confident his bank can quickly regain its position.”

In the above example, 'he' and 'his' both refers to the 'Sumitomo President Koh Komatsu' were correctly identified by the personal pronoun filter.

3.2. Salience weights

After applying all above mentioned filters, there is a possibility that for a particular anaphora more than one potential antecedent are present. Hence, in order to get final antecedent salience weight is used. [2]

The main text format consists of a flat left-right columns on A4 paper (quarto). The margin text from the left and top are 2.5cm, right and bottom are 2 cm. The manuscript is written in Microsoft Word, single space, Time New Roman 10pt and maximum 12 pages, which can be downloaded at the website: http://www.iaesjournal.com/online/index.php/IJECE

A title of article should be the fewest possible words that accurately describe the content of the paper. Omit all waste words such as "*A study of ...*", "*Investigations of ...*", "*Implementation of ...*”, "*Observations on ...*", "*Effect of.....*", “*Analysis of …*”, “Design of…” etc. Indexing and abstracting services depend on the accuracy of the title, extracting from it keywords useful in cross-referencing and computer searching. An improperly titled paper may never reach the audience for which it was intended, so be specific.

The Introduction should provide a clear background, a clear statement of the problem, the relevant literature on the subject, the proposed approach or solution, and the new value of research which it is innovation. It should be understandable to colleagues from a broad range of scientific disciplines. Organization and citation of the bibliography are made in Vancouver style in sign [1], [2] and so on. The terms in foreign languages are written italic (italic). The text should be divided into sections, each with a separate heading and numbered consecutively. The section/subsection headings should be typed on a separate line, e.g., **1. Introduction** [3]. Authors are suggested to present their articles in the section structure: **Introduction - the comprehensive theoretical basis and/or the Proposed Method/Algorithm - Research Method - Results and Discussion – Conclusion**.

Literature review that has been done author used in the chapter "Introduction" to explain the difference of the manuscript with other papers, that it is innovative, it are used in the chapter "Research Method" to describe the step of research and used in the chapter "Results and Discussion" to support the analysis of the results [2]. If the manuscript was written really have high originality, which proposed a new method or algorithm, the additional chapter after the "Introduction" chapter and before the "Research Method" chapter can be added to explain briefly the theory and/or the proposed method/algorithm [4].

4. SYSTEM ARCHITECTURE

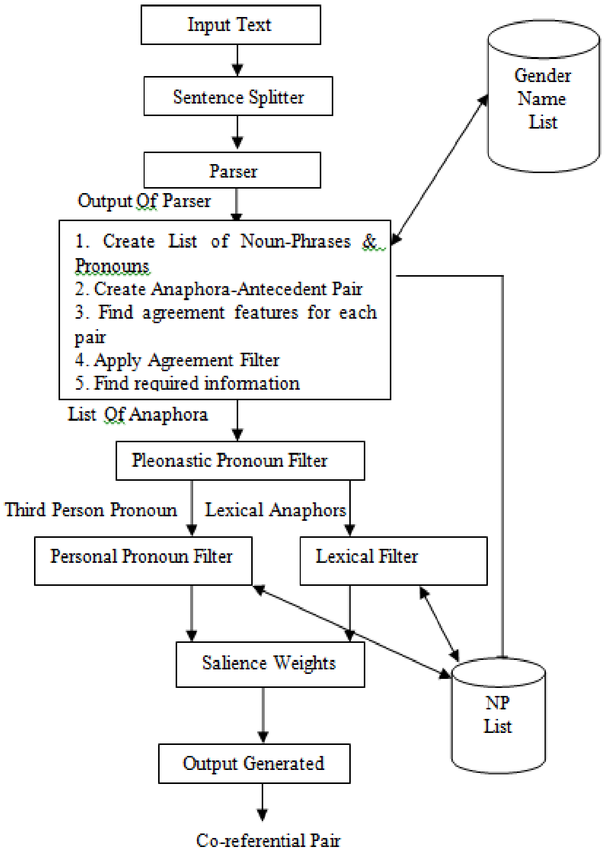

Figure 2. Architecture of "Automated Anaphora Resolution".

Instead of using a single monolithic architecture system uses the hybrid approach which combines constraint-based and preferences-based architectures, as shown in the Figure 1. System takes input in the form of text files and assigns control to the sentence splitter. Sentence splitter spits the sentences and assigns tags in the beginning and end of each sentence as per required by the parser. Output of the sentence splitter is then given to the parser. Parser used by the system is Charniak's Parser (parser05Aug16) which tags the text and generates parse tree. From the syntactic structural generated by the parser, system generates two lists of anaphora and noun phrases. From the list of pronoun and noun phrases, all possible pairs of anaphora and antecedents are generates. Each pair is then filtered through agreement filter which checks for compatibility of each pair on the basis of agreement features. All the information required by the system is not generated by the parser; hence remaining required information is evaluated by the system for the further processing. Next step is to apply pleonastic pronoun filter which will take list of anaphora as an input. After applying pleonastic filter, personal pronoun filter is applied; it considers the list of anaphora and list of noun phrase created by the system. There is a possibility that for a particular anaphora more than one potential candidate antecedent can be present. So, from the list of all potential candidate antecedents, final antecedent is chosen with the help of salience weight. The architecture of "Pronominal Anaphora Resolution" is given in Figure 1 in the pictorial form above.

5. ISSUES IN ANAPHORA RESOLUTION

The Basically, there are two main approaches for resolving anaphora: (1)Traditional Approach, which usually depends upon linguistics knowledge, and (2) the Discourse-oriented Approach, here the researcher tries to model complex discourse structure and then uses structures for the process of anaphora resolution. Traditional approaches apply linguistics knowledge, in the way of “Preferences” and “Constraints”, in which systems can be proposed as a technique for combining various information sources. Traditional approaches are mainly works in basic three steps as: (1) Deciding search limit or anaphoric accessibility space, (2) apply various constraints, and then (3) apply preferences. [3]

A. Search limit or Anaphoric Accessibility Space:

This represents a limit for searching all possible candidate antecedents for a particular anaphora. System defines the text segments within which antecedent for a particular anaphora can be found. Finding 'Anaphoric Accessibility Space' is a very crucial phase in the process of anaphora resolution. Because, small search limit results in the exclusion of valid antecedents and too broad search limit results in large candidate lists, which ultimately results in erroneous anaphora resolution. Generally, search limit or 'Anaphoric Accessibility Space' is defined as 'n' previous sentence to anaphora, where value of 'n' varies according to the kind of anaphora. According to Ruslan Mitkov (2008), the ideal anaphora resolution system, value of 'n' is 17 i.e. System check 17 sentences away from the sentence in which anaphora is present. 'Anaphoric Accessibility Space' is predefined by the developer, if any anaphoric word is recognized by the system then list of all possible candidates for antecedent within the predefined search limit is found out for that particular anaphora.

B. Constraints

After getting the list of all possible antecedents, several constraints are applied for removing the incompatible antecedent. Usually, constraints hold certain conditions that must be fulfilled by the candidate antecedent, if it fails, then candidate will not be considered possible antecedents for the anaphor. Various information like syntactical, semantic, morphological, and lexical are used to define various constraints.

C. Preferences

After removing all incompatible candidates for antecedent, if the remaining list of antecedent contains more than one antecedent, then preferences are applied for selecting only one potential antecedent. Preference system must be developing by keeping in mind that only a single candidate must remain at the end of process. And these final candidates given by the preference system will be considered as a potential candidate for that particular anaphora. Various information like syntactical, semantic, morphological, and lexical are used to define preference system.

A text contains linguistic data in many form such as syntactic parallelism, antecedent proximity, gender and number agreement, lexical repetition or c-command restrictions plays an vital role in the anaphora resolution process. Various methodologies such as statistical and probabilistic models, Knowledge-poor solutions, using corpus-driven methodologies are preferred for resolving anaphora. Opposite to the pure statistical model, some strategical approaches have also been proposed for tracking the antecedent, which can be formalized in terms of rules based on 'preferences' and 'constraints'. Such strategical approach usually combines of 'constraints' and 'preferences'.

The working of anaphora resolution system relies on the set of various anaphora resolution factors. These factors can be either “eliminating” i.e. it does not count some candidates from the list of all possible candidates or “preferential” i.e. It give more preference to some candidates than remaining candidates. Partition of anaphora resolution factors into preferences and constraints is responsible for the preferences-based and constraints-based architecture in anaphora resolution. Instead of using single preferences-based or constraints-based architecture, in our system we have use the combination of preferences-based and constraints-based architecture, in order to get the more efficient results. We have used this architecture for our system because study shows that anaphora resolution systems based on constraints and preferences can give a successful result when applied to non-dialogue texts.

6. EVALUATION

Testing of System performs manually. For the testing purpose, we have use 'Reuters Newspaper corpus'. In Reuters Newspaper corpus', total 4024 files are present in the test section and total 11,413 files are present in the Training section. In 'Reuters Newspaper corpus', files are present in total 91 various categories like housing, income, jobs, money-supply, livestock, retail, interest, silver, trade and so on in both test and training section. In the 'Reuters Newspaper corpus', total 15,437 files are present distributed over 91 different categories in both test and training section. Files in the 'Reuters Newspaper corpus' are present in TXT format, of minimum size 693 byte and maximum size of 4.3 kb approximately.

‘Table III’ shows that testing of system is performed on total 120 file of different categories of 'Reuters Newspaper corpus'. In these 120 files, total 492 anaphora-antecedent pairs were found out of which 404 pairs were correctly identified by the system. Hence, we can say that efficiency of the system is 82.11%. 120 files which we have used for the testing purpose are present in TXT format, of minimum size 693 byte and maximum size of 4.3 kb approximately. We can say that the average size of each file is 2.5 kb. On an average, we can say that each file contains approximately 4 pairs of anaphora-antecedent pair.

Table 3. Shows that Testing of System

Total File considered	Total Number Of Anaphora-antecedent Pairs Present	Number of cases that the system resolves correctly	Accuracy %
120	492	404	82.11 %

In total 492 pairs of Anaphora-antecedent, all third person personal (he, him, she, her, it, they, them), possessive (his, her, hers, its, their, theirs), and reflexive (himself, herself, itself, themselves) pronouns in both singular and plural, are tested along with the occurrence of pleonastic it and the lexical anaphora. From the result coming out from the experiment, system correctly founds total 27 pairs of third person personal from 45 pairs , 144 pairs of possessive pronouns from 167 pairs, and 191 occurrence of ‘It’ pronouns from 230 pairs in both singular and plural form, also 42 pairs of lexical pronouns from 50 pairs as shows in figure 3. Hence we can say that system have 60% efficiency to find third personal, 86.22% efficiency to find possessive pronoun, 83.04% efficiency in finding ‘It’ pronoun, and system have 84% efficiency to find lexical anaphora.

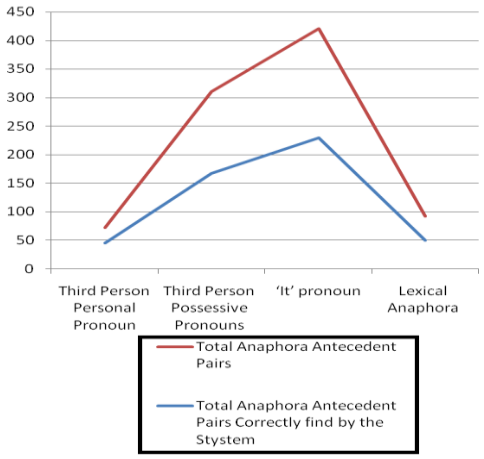

Figure 3. Performance of System

We can say that system almost founds 27 of third person personal, 144 pairs of possessive, 191 pairs of ‘it’ pronouns and 42 pairs of lexical pronoun. In order to improve the performance of the system powerful personal pronoun and agreement filter plays an important role. The way in which implementation extract the grammatical roles by applying certain hand-crafted rules on the parse tree also affect the overall performance of the system. Use of knowledge rich Charnak parser (parser 05 aug 16) in system, also contributes in overall performance of the system.

Along with the intra-sentential anaphora system also finds inter-sentential anaphora. Out of the total 492 pairs of anaphora and antecedent, total 387 pairs were intra-sentential and total 105 pairs were found to be inter-sentential. Out of total 387 intra-sentential pairs, system founds 314 anaphora-antecedent pair correctly and out of total 105 inter-sentential pair system finds 90 pairs correctly. Hence, efficiency of system to find inter-sentential and intra-sentential anaphora-antecedent pairs is 85.71 and 81.13 respectively, which is shown by Figure 4 in a pictorial form.

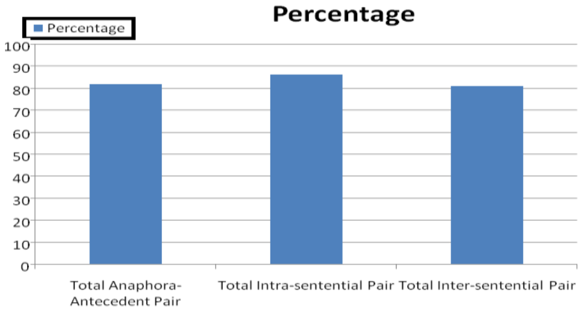

Figure 4. Efficiency of System in Percentage

7. CONCLUSION

System which uses the combination of constraint-based and preferences-based architectures, for resolving anaphors is represented as above. It is used to identify inter-sentential and intra-sentential antecedents of “Third person pronoun anaphors”, “Pleonastic it”, and “Lexical noun phrase anaphora”. An algorithm at first defines an anaphoric accessibility space, then applying constraints, and finally applying preferences. Anaphoric accessibility space is selected carefully which is not too broad or too small. Several constraints are applied in order to remove incompatible antecedents for a particular anaphor. Finally, after removing incompatible candidates, if the remaining list contains more than one antecedent, salience weights are applied in order to choose a single antecedent. The algorithm applies to the output generated by Charniak parser (parser05Aug16) and relies on salience measures derived from parse tree. Various factors such as "eliminatory" i.e. doesn’t count certain noun phrases from the set of potential antecedents (such as gender , number , people constraints) and "preferential" i.e. assigning more preference to some potential antecedents and less to others (such as salience) is used in order to resolve the anaphora. We have tested the system extensively on 'Reuters Newspaper corpus' and efficiency of the system is found to be 82.11%.

REFERENCES

[1] Ruslan Mitkov, *ANAPHORA RESOLUTION: THE STATE OF THE ART*, International Conference on Mathematical Linguistics, 2008

[2] Shalom Lappin and Herbert J. Leass ,An Algorithm for Pronominal Anaphora Resolution, 1994

[3] Chinatsu Aone and Scott William Bennett, *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies,* International Workshop on Sharable Natural Language Resources (SNLR),2000

[4] Mitkov and Ruslan, Anaphora resolution in Natural Language Processing and Machine Translation. Working paper. Saarbrücken: IAI, 1995a.

[5] Mitkov, Ruslan, *"Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches"* Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution. 14-21. Madrid, Spain, 1997b.