An Extensive Evaluation of Anaphora Resolution Based Abstract Generation System

Ayyalu Hariharan Nandhu Kishore
Department of Computer Science and Engineering
Madurai Institute of Engineering and Technology (MIET)
Madurai, Tamil Nadu, India.
ahnandhume@yahoo.co.in

Mohan Saravanan
Department of Computer Science and Engineering
Chennai, Tamil Nadu, India
msdess@gmail.com

***Abstract*— This paper presents an extensive evaluation of anaphora resolution (AR) algorithm proposed for Abstract Generation System (AGS) and discusses the metrics used to measure its performance. The evaluation exercise is conducted with the test sets chosen from different domains and the results are compared with other existing open-source Anaphora Resolution Systems (ARSs) like Mitkov's Anaphora Resolution System (MARS) and Java Resolution of Anaphora Procedure (Java RAP). We have selected news documents belong to five different domains and it is trained with AGS among which some set of documents and their corresponding extract is taken as the test set and executed in various ARSs for comparison. The proposed system is evaluated and its performances are studied with standard metrics of Information Extraction (IE) systems such as Success rate, Precision, Recall and F-measure.**

Keywords- Abstract Summary; Dangling Anaphora; Natural Language Processing (NLP); Text-mining Framework; IE Metrics

I. INTRODUCTION

The amount of data available online grows significantly day by day. Even for a single search returns huge volume of web pages which are very difficult for the internet users to read and understand completely. In other words, a good intelligent *Information Extraction (IE) tool* summarizes all available related information in the form of a summary is a valuable solution for this purpose (e.g. *automatic text summarization*). A summary is defined as the collection of text that is produced from one or more texts that contain a significant portion of the information in the original text and that is no longer than half of the original text [1]. When the summary is generated automatically with the help of computer, then it is referred to as Automatic Text Summarization [2]. The output produced by the automatic summarizer can considered as an extract summary that is selecting important sentences of a document based on the application of statistics or semantics. The percentage level of summarization is fixed and the details are provided to the summarizer. Other form of summarization is based on abstraction, which is generated newly with the extracted details from the original document. It is a well known fact that it is highly difficult to generate abstract summary as a fully synthesized text for a given document. We have suggested a novel way that it is possible to generate an abstract summary by fine-tuning the extract summary after applying an *anaphora resolution algorithm* on the extract and the original text.

The term anaphora in anaphora resolution refers to the linguistic phenomenon of referring back to a previously mentioned entity in the text, most often with the help of a pronoun or a different name. Anaphora resolution (AR) is the process of resolving the anaphors with its correct antecedent within the context [3]. The interpretation and resolving an anaphor with its correct antecedent is one of the outstanding issues and hot research areas in the computational linguistics [4]. The application of anaphora resolution includes the text summarization, question-answering systems, location tracking system etc.

When the anaphora is a pronoun within the context, then it is called as *pronominal anaphora* [4]. The process of resolving the pronominal anaphors is called as pronominal anaphora resolution. The algorithm which is designed to perform this process is called *pronominal anaphora resolution algorithm* [5] [6] [7].

As said earlier, to generate an abstract, pronominal anaphora resolution algorithm can be applied [8] which should resolve the pronouns with their correct antecedent, reflect it back in the extract. It also process the broken anaphoric reference that occurs due to the disconnection between the sentences while extracting the summary. The above said process is referred to as *dangling anaphora resolution*. The anaphora resolution based abstract generation system should perform both the pronominal anaphora resolution and the dangling anaphora

resolution so as to result in improvement of the extract summary and obtain a coherent and readable abstract summary [9] [10] [11].

The literature survey reveals that there exist two major approaches of anaphora resolution. One is knowledge based approach and the other one is the knowledge poor approach. The knowledge based approach is a labor-intensive task which employs the parser for analyzing and resolving the anaphors and also takes syntactic and semantic knowledge of each and every word in the document. The public reference implementation of anaphora resolution namely Java RAP is designed based on this approach [12] [13].

The knowledge poor approach neither uses the parser nor the syntactic or semantic knowledge, rather than it makes use of a Parts of Speech (POS) tagger and noun phrase rules to resolve the anaphors. The online anaphora resolution engine namely Mitkov's Anaphora Resolution Systems (MARS) [14] [15] is implemented under this strategy. We also followed the same strategy in our approach of anaphora resolution system.

This paper presents the extensive evaluation of anaphora resolution algorithm proposed for AGS with the help of IE metrics such as success rate, precision, recall and F-measure. F-measure is defined as a weighted harmonic mean of precision and recall measures. In Section 2, we explain the structure of AR based AGS in detail. Section 3 describes the metrics and measures that are used for the comprehensive evaluation of anaphora resolution algorithm. Section 4 discusses the evaluation results of the algorithm where as Section 5 concludes this paper and discuss the further work in this study.

II. ANAPHORA RESOLUTION BASED ABSTRACT GENERATION MODEL

Figure 1 illustrates the overview of AR based abstract generation system. The source document is the original document for which the abstract summary has to be generated. The source document and its extracted summary form the input to the anaphora resolution based abstract generation model. The extract form of the summary is generated with the help of online automatic summarization method like MEAD [16].

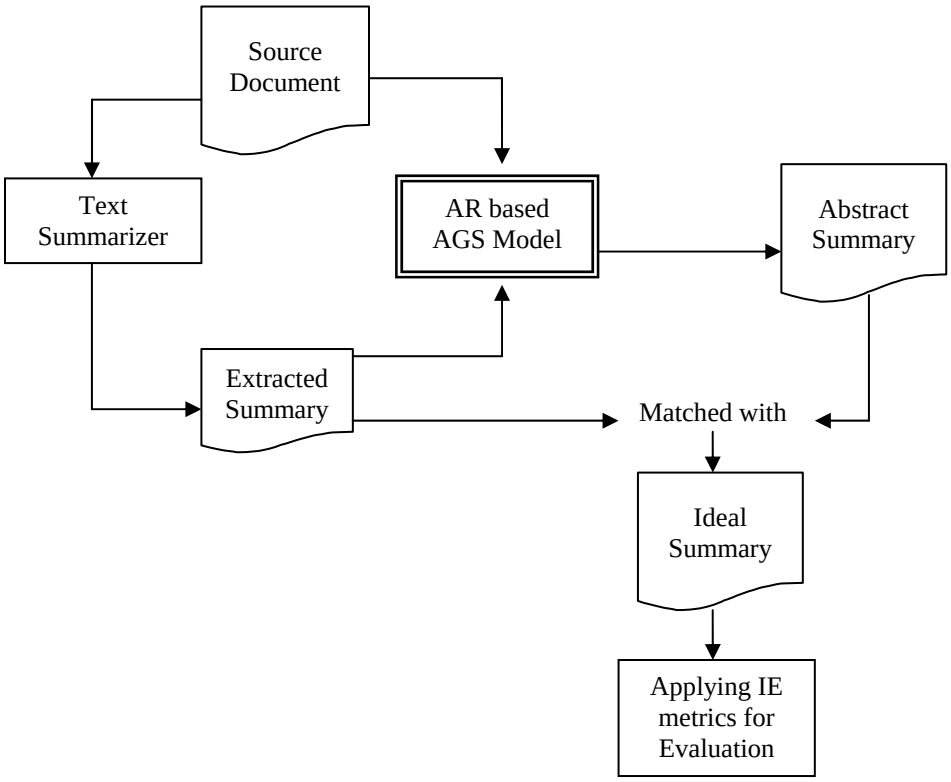


Figure 1. Overview of Anaphora Resolution Based Abstract Generation System

The output of this model is a smoothened summary. This improved summary will be compared with ideal summary generated by human judges. The performance is evaluated using various IE metrics and the details are provided in Section 3.

A. Structure of Abstract Generation Model

Figure 2 shows the text-mining framework of abstract generation model including the sub-modules in it. The abstract generation process involves three steps. They are the generation of extract summary, preprocessing the input document and then finally the abstract generation phase.

The extract summary is obtained by executing the source documents in the automatic text summarizer namely the MEAD summarizer. It is a public domain portable multi-document summarization system. It provides the summaries for the documents on different percentage levels.

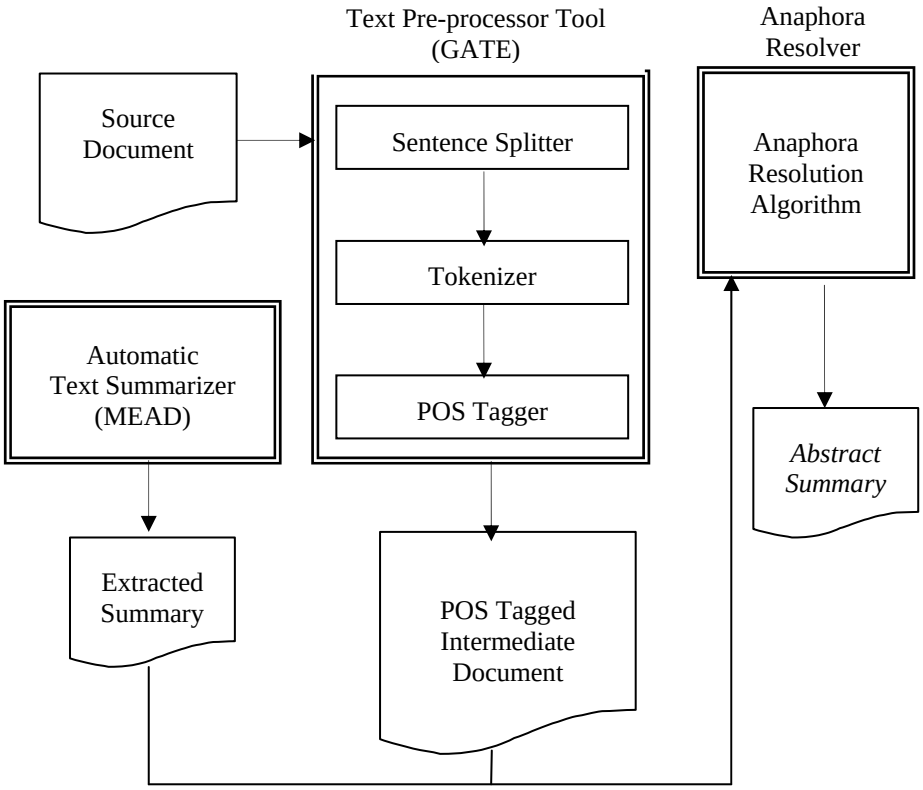


Figure 2. Structure of Abstract Generation Model

The text preprocessing phase is carried out to make the process of resolving the anaphors easier. The text pre-processor tool used in our system is the General Architecture for Text Engineering (GATE) [17]. It consisting of three modules executed in a pipeline. They are the sentence splitter plug-in that is used to split the paragraphs in the document into individual sentences, followed by the Tokenizer plug-in, which will divide sentences into individual words. Finally the POS tagger plug-in provides the POS tag to individual words. The GATE tool is trained to identify the gender of the named entities. This POS tag is used to identify the gender of the antecedent which in turn paves th way to resolve the anaphors for the pronouns. The output of this text pre-processor is saved as POS tagged file.

Any pronoun present in the document matches within the considered pronoun set, then the corresponding antecedent is fetched from the source document and it is replaced in the extract summary. The system is designed as a semi-automatic system which requires human intervention for generating the extract and preprocessing the document with separate interface. The document and its corresponding extracted summary is fed as input to the AR incorporated with AGS (as shown in Figure 1) which results in the generation of abstract summary after resolving and fixing the anaphors for the antecedents making the summary more readable. The category of pronouns considered for resolution includes the personal pronouns, non-anaphoric 'it' (pleonastic 'it') and dangling anaphors (in extract). The algorithm can be easily extended for other pronouns also.

III. EVALUATION METRICS AND MEASURES OF AR ALGORITHM

The main theme of the paper is to present the evaluation metrics and measures of AR algorithm used in abstract generation system. The Information Extraction metrics namely the success rate, precision, recall and F-measure [18] [19] [20] are used to evaluate our algorithm. These metrics are defined as follows.

$$\text{Success (S)} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors.}} \quad (\mathbf{1})$$

$$\text{Precision (P)} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of anaphors attempted to be resolved.}} \quad (2)$$

$$\text{Recall (R)} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors to be resolved}} \quad (3)$$

$$\text{F-Measure} = \frac{2 * P * R}{(P + R)} \quad (4)$$

To calculate the above said measures, the following quantities are needed.

- The total number of anaphors present in the documents
- Anaphors identified by the system
- Correctly resolved anaphors

The personal pronouns including *he*, *she* and *it*, is alone considered at present for resolution because here the anaphora resolution is used as a tool to generate the abstract summary. If not applied in AGS, it can easily be extended to resolve any kind of anaphor present in the text.

The online news documents about 100 in numbers belonging to five different domains are selected which is described in Table 1, are downloaded and the extract summary for those documents are generated. The source documents were trained in GATE for named-entities. From the trained corpus, a test case set containing 50 documents and their corresponding extracts are selected and executed in our system. The test case corpus contains a total of 25,734 words with 325 pronouns out of which 264 are anaphoric pronouns. The total numbers of correctly resolved anaphors are 229. The result of the execution of anaphora resolution algorithm on the test documents appears in Table 1.

TABLE I. RESULTS OF ANAPHORA RESOLUTION ALGORITHM IN AGS

Domain	He		She		It		Dangling anaphors		Pleonastic it	
	I	*II*	*I*	*II*	*I*	*II*	*I*	*II*	*I*	*II*
Sports	14	12	13	13	10	9	8	7	12	11
Politics	15	15	10	10	8	5	4	3	16	15
Education	14	13	13	10	11	10	6	5	11	8
Arts	13	10	13	11	9	9	5	5	10	9
Business	17	12	10	9	5	4	5	4	12	10
Total	**73**	**62**	**59**	**53**	**43**	**37**	**28**	**24**	**61**	**53**

In the above table, column I refer to total number of corresponding anaphors present in the document and the column II represent the total number of correctly resolved anaphors by the system. It is inferred from the table that some of the anaphors '*He*' and '*She*' are not correctly resolved. This error mainly occurs to the selection and fixing of the recent antecedent found at that context which actually refers to some other antecedent. This is also due to ambiguities in antecedents that occur in situations. Otherwise the system works well and resolves the anaphors correctly.

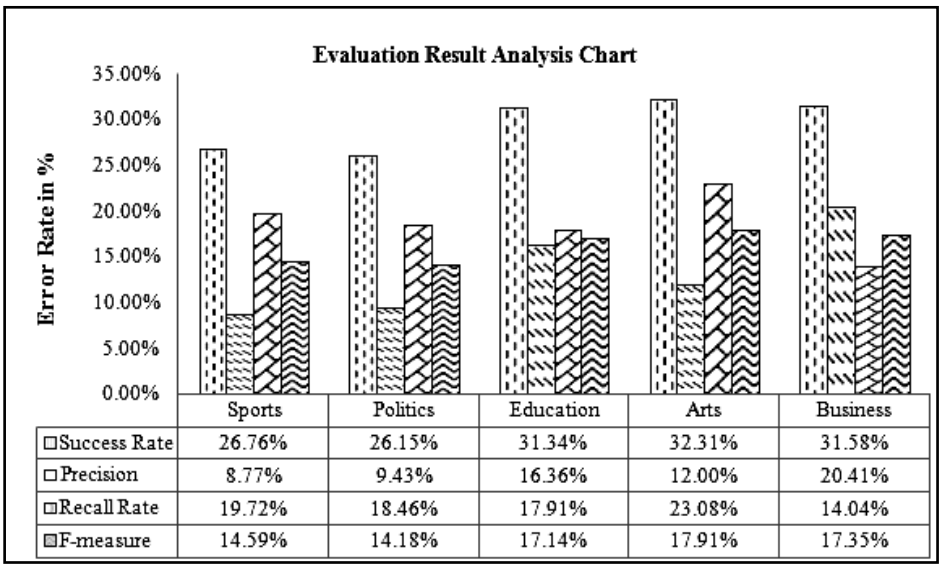
TABLE II. EVALUATION RESULTS OF ANAPHORA RESOLUTION BASED AGS

Domain	Success Rate	Precision	Recall Rate	F-measure
Sports	73.24%	91.23%	80.28%	85.41%
Politics	73.85%	90.57%	81.54%	85.82%
Education	68.66%	83.64%	82.09%	82.86%
Arts	67.69%	88%	76.92%	82.09%
Business	68.42%	79.59%	85.96%	82.65%
Average	**70.37%**	**86.61%**	**81.36%**	**83.77%**

On the other hand, when a name of person entity or object has two or more words it will consider only the last word as antecedent and it will be replaced with the anaphor. This is because the system is designed under knowledge poor approach and doesn't consider the semantics.

The evaluation results of the above said measures are presented in Table 2. From this table, it is clear that the success rate of the system has been gone down for the different domain considered. The reason may be that in our system the anaphora resolution is applied only to a set of pronouns and not to all kinds. In contrary to that we expect our system has to generate the abstract summary more coherent and readable without affecting the meaning to be expressed in it. Hence only the personal pronouns are considered.

The performance chart of the system shown in Figure 3 is plotted using the error rate values of IE metrics thus obtained over various domains.



	Sports	Politics	Education	Arts	Business
Success Rate	26.76%	26.15%	31.34%	32.31%	31.58%
Precision	8.77%	9.43%	16.36%	12.00%	20.41%
Recall Rate	19.72%	18.46%	17.91%	23.08%	14.04%
F-measure	14.59%	14.18%	17.14%	17.91%	17.35%

Figure 3. Error Rate Evaluation Analysis Chart of AR based Abstract Generation System

From the above bar chart, it is obvious that the error rate of precision is very much less comparatively which shows that the system has resolved most of the anaphors that are attempted for resolution with very correct antecedent. For confirmation, the output of the system, the abstract summary is checked with its ideal summary. It is because in many cases, the recent antecedent will be the right antecedent. Our system too does this well. On an average, the system has achieved 87% in precision scoring.

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

IV. RESULTS AND DISCUSSIONS

In the pronominal AR evaluation experiment, the following results were obtained.

- The types of pronouns vary considerably according to the domain of the test corpus.
- It contains a lot of person entities that are referred in the text with the help of personal pronouns like he and she. The rest of the references are the object entities using ‘it’ pronouns.
- The reason for most of the anaphors has been got resolved correctly is due to
    - Training the documents using the GATE tool for named-entities. The male names, female names and objects are differentiated using special user defined tags which can be provided to a file namely lexicon in the GATE software before executing it. It is enough to enter this special tag only once for a particular entity. When any document encounters that entity, the special tagwill be assigned to that antecedent while training the document. This is the breakthrough in our research which helped us to achieve a considerable increase in the performance of the system..
    - Many of the papers have defined the term ‘dangling anaphora resolution’ but they lacked in describing the clear idea of its implementation. In our system it has been accomplished and executed successfully. Among twenty eight dangling anaphors only four have been mistakenly resolved. This also added the improvement in the performance of the system.
- The reason for certain anaphors that is not resolved correctly is originated from various causes.

- The system considers the POS tag for each and every word. Based on that tag, it resolves the anaphors with recent antecedent available. It does not consider the semantic information. This is one of the main reasons to lower the success rate of our system.
- There were twenty three occurrences of *he*, *she* and *it* pronouns that are resolved incorrectly due to selecting the recently found antecedent for those anaphors.
- On the other hand, there were eight mistakenly resolved '*it*' pronouns. This is because the system has failed to differentiate the anaphoric '*it*' with pleonastic '*it*'. Also '*it*' pronouns lacks gender information.
- It is a slightly difficult task to differentiate the anaphoric '*it*' and pleonastic '*it*'. However, our system tries to differentiate significantly by considering the case that when an antecedent for '*it*' is present in previous sentence, then it is considered as anaphoric *it*. Otherwise it is taken as non-anaphoric *it*.

V. CONCLUSION AND FUTURE ENHANCEMENTS

The evaluation of the system has been carried out with the trained corpus. The efficiency of the system has been achieved to a sensible level of about 87%. Our system performance is comparatively better than other systems considered in this study. Since our system depends on the external resource for extract generation and text pre-processing with separate interface, the test results may not compared with the online anaphora resolution systems like MARS and Java RAP which are designed mainly for performing anaphora resolution. Our system is designed for performing the task of text-summarization. We have a plan to modify the system to consider the semantic information of words into account and also to check the bi-grams and tri-grams to resolve an anaphor with the name of a person or object containing more than two or more words. Word Net, a lexical database plug-in can be included in the text pre-processor tool to consider the semantic information of the words. It can also be extended to synthesize multi-document abstracts by running it in a fully automatic mode by integrating all the interfaces into a single system.

REFERENCES

[1] Elena Lloret, Oscar Ferrandez, Rafad Munoz, Manuel Palomer: A Text Summarization Approach under the Influence of Texual Entailment, In: The 5th International Workshop on NLP and Cognitive Science (NLPCS 2008) in conjunction with the 10th International Conference of Enterprise Informaation System (ICEIS 2008), pp. 22 –31, Barcelona, Spain , 2008.

[2] Eduart Hovy and Chin-Yew Lin: Automatic Text Summarization in SUMMARIST In: Advances in Automated Text Summarization, MIT Press, pp. 18--24, USA, 1998.

[3] Allaoua Refoufi: A Modular Architecture for Anaphora Resolution. In: The Journal of Computer Science, Vol. No. 3 (4), pp. 199–203, April , 2007.

[4] Ruslan Mitkov: Outstanding issues in Anaphora Resolution. In: Second International Conference on Computational Linguistics and Intelligent Text Processing, pp. 110–120, Mexico, February 18 – 24, 2001

[5] Anna Kupsc, Teruko Mitamura, Benjamin Van Durme, Eric Nyberg: Pronominal Anaphora Resolution for Unrestricted Text. In: International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, pp 1495--1498, May 24 – 30, 2004.

[6] Marin Dimitrov, Kalina Bontcheva, Hamish Cunningham, Diana Maynard: A Light – Weight Approach to Coreference Resolution for Named Entities in text. In: MSc Thesis, University of Sofia, Bulgaria, 2002.

[7] Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, Karel Jeek: Two Uses of Anaphora Resolution in Summarization, Information Processing and Management: an International Journal, Vol 43 (6), pp 1663–1680, Pergamon Press, Inc. Tarrytown, NY, USA, 2007.

[8] Saravanan.M, Reghu Raj P.C and Raman.S: Summarization and Categorization of Text Data in High Level Data Cleaning for Information Retrieval, In: First International Workshop on Data Cleaning and Preprocessing, pp. 119--130 , Maebashi City, Japan, December 9 – 12 , 2002.

[9] Patricia Nunes Goncalves, Lucia Rino, Renata Viera: Summarizing and referring: towards cohesive extracts, In: Eighth ACM Symposium on Document Engineering, pp. 235--236, Sao Paulo, Brazil, September 16 – 19, 2008.

[10] Nandhu Kishore A.H, Saravanan.M: Anaphoric Resolution for Extracted Summaries on Different domains. In: International Conference on Advanced Computing and Communication (ICACC' 2007), pp. 201--204, Madurai, Tamil Nadu, India, February 9 – 10, 2007.

[11] Nandhu Kishore A.H., Saravanan.M: A Novel Text-Mining System for Generating Abstract from Extracted Summaries Using Anaphora Resolution, In: International Conference in Data Engineering and Management (ICDEM 2010) in Post Conference Proceedings in Lecture Notes in Computer Science (LNCS), Springer-Verlag, Germany, Vol.No. 6411, pp. 36 – 43, 2011.

[12] Long Qiu, Min-Yen Kan, Tat–Seng Chua: A public Reference Implementation of the RAP Anaphora Resolution Algorithm, In: Language Resources and Evaluation Conference 2004 (LREC 04), Lisbon, Portugal, May 26–28, 2004.

[13] The Online Demo for Java RAP : http://www-appn.comp.nus.edu.sg/%7Erpnlpir/cgi-bin/JavaRAP/JavaRAPdemo.html

[14] Ruslan Mitkov: Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In Discourse, Anaphora and Reference Resolution Conference (DAARC2000), pp 96--107, Lancaster, United Kingdom, 2000.

[15] The Online MARS Engine: http://clg.wlv.ac.uk/demos/MARS/index.php

[16] Online Demo of MEAD Summarizer: www.newsinessence.com/clair/meaddemo/demo.cgi
http://www.summarization.com/mead

[17] GATE URL: http://gate.ac.uk/sale/tao/

[18] Rodolfo Delmonte, Antonella Bristot, Marco Aldo Piccolino Boniforti, Sara Tonelli: Another Evaluation of Anaphora Resolution Algorrithms and a Comparison with GETARUNS's Knowledge Rich Approach, In ROMAND 2006 – 11th EACL, pp. 3—10, Geneva,

2006.

[19] MijailnAlexandrov Kabadjov: A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification, Ph.D thesis, University of Essex , 2007.

[20] Barbu, Catalina, Ruslan Mitkov: Evaluation Tool for Rule-Based Anaphora Resolution Methods, In Proceedings of ACL 2001, pp. 34-41, 2001.

AUTHORS PROFILE

A.H.Nandhu Kishore has received his Bachelor's Degree in Computer Science and Engineering from K.L.N. College of Engineering, (MKU University), Madurai, Tamil Nadu, India in 2004. He received his Master's Degree in the same discipline from Jerusalem College of Engineering (Anna University), Chennai, India.

He is serving as Assistant Professor and HOD in Department of Computer Science and Engineering in Madurai Institute of Engineering and Technology (MIET), Madurai, Tamil Nadu, India.

M.Saravanan received MS and Ph.D degrees in Computer Science & Engineering from Indian Institute of Technology Madras (IITM), India. His research work has a significant contribution towards Legal Information Retrieval.

His research interests include Probabilistic models for machine learning, Text data mining, Statistical Natural Language Processing and their applications to real world problems.