INFORMATION EXTRACTION USING DISCOURSE ANALYSIS FROM NEWSWIRES

Ashwini Rahangdale[1] and Dr.A.J.Agrawal[2]

[1]M.Tech Scholar,Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering & Management, RTM University, Nagpur,India
[2]Associate Professor, Department of Computer Science, Shri Ramdeobaba College of Engineering & Management Nagpur, India

ABSTRACT

This paper proposes Natural language based Discourse Analysis method used for extracting information from the news article of different domain. The Discourse analysis used the Rhetorical Structure theory which is used to find coherent group of text which are most prominent for extracting information from text. RST theory used the Nucleus- Satellite concept for finding most prominent text from the text document. After Discourse analysis the text analysis has been done for extracting domain related object and relates this object. For extracting the information knowledge based system has been used which consist of domain dictionary .The domain dictionary has a bag of words for domain. The system is evaluated according gold-of-art analysis and human decision for extracted information.

KEYWORDS

Discourse parser, Rhetorical Structure Theory, Elementary Discourse Unit, Nucleus, Information Extraction.

1. INTRODUCTION

Natural Language Processing is a theoretically motivated range of computational linguistic for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. There are different levels like semantic analysis, opinion analysis, text summarization, information extraction, information retrieval that should be studied in order to understand this computational linguistic.

Discourse can be defined as language beyond the level of sentence or Language behaviors linked to social practices or language as a system of thought. Discourse Analysis (DA) is a modern discipline of the social sciences that covers a wide variety of different sociolinguistic approaches. It aims to study and analyse the use of discourse in at least one of the three ways stated above, and more often than not, all of them at once. Analysis of discourse looks not only at the basic level of what is said, but takes into consideration the surrounding social and historical contexts. A discipline of DA is 'Discourse Analysis' looks at discourse from a politically motivated level. An analyst in this field will identify a topic for analysis, and then collect a corpus of texts, before finally analysing it to identify how language is used to reproduce ideologies in the text. A corpus is large, structured electronic database of texts, often used in linguistics. Using a corpus isn't the only method of analysis in DA, as any method which provides an insight into ideology in discourse is accepted by researchers.

In a well-written text, no unit of the text is completely isolated; interpretation requires understanding the unit's relation with the context. Research in discourse analysis aims to unmask such relations in text, which is helpful for many downstream applications such as summarization, information retrieval, and question answering.

An information extraction (IE) system analyses unrestricted, real world text such as newswire stories. In contrast to information retrieval systems which return a pointer to the entire document, an IE system returns a structured representation of just the information from within the text that is relevant to a user's needs, ignoring irrelevant information. Information extraction using discourse analysis is divided into two steps first is the Discourse analysis and second is content analysis. In first stage of Discourse analysis merges together multiple references to the same objects identify the logical relationship between the different sentence and finding the most prominent sentence part and infers information not explicitly defined by the sentence analysis. In the second, discourse sentence analysis has been done for content selection identifying the relevant object for the discourse identification and typically creates the case frame for representing referenced object. IE system operates on domain specification that predefines what types of information are considered relevant to the application. Domain Knowledge is used for referencing domain object, creating relationship between the different object and description of object according to the domain.

Newspaper or News web-site contains the day to day news of the different domain like sports, businesses, politics etc. This domain can be identified using the discourse analysis and text analysis. After the identification of domain abstract of news can be extracted from news document.

Remaining of this paper organized as: section 2 explains motivation behind the discourse analysis for extracting information from news article. Section 3 explains related work till date

2. RELATED WORK

In the Information Extraction using discourse analysis I have consider two different tasks combining together. In domain worked independently.

Researcher in linguistic and computational linguistic s have long pointed out that text is not just simple sequence of clauses and sentence, but rather follows a highly elaborated structure. One of the approaches to deal with linguistic is discourse analysis. Discourse analysis used for text based application like information extraction, dialogue generation, and summarization. The framework used for Discourse analysis is Rhetorical Structure Theory (RST). Rhetorical structure theory proposed by Daniel Marcuu (1988) [5,6].This framework has used for identification of discourse marker in the number of natural language text dealing with single subject matter, Marcu(8,10) has shown that RST can be used for automated mark-up of natural text and has shown the identification of discourse marker from prototypical text can be automated with 88% precision compare to those identified perfectly by human analysis RST theory provides framework to analyses and study text coherence by identifying and applying a set of structural relation to composing units (span) of text [1,3].

According to Mann and Thomson (1988)[14] all well written text is supported by hierarchically structured set of coherence relation which reflects author's intent. The goal of discourse parsing provides information on grammatical structure of text. Discourse parsing and other higher-level view of text allowing some flexibility in the choice of formal representation. This tree likes structure has been used by (piwek et al, 2007) [15] for text generation. The tree used by Daniel marcuu (2000) [6] for extracting summary from text. Micheal Regneri and Rai wong proposed a

novel method for collecting paraphrases relying on the sequential event order in the discourse, using multiple sequence alignment with semantic similarity measure. They have shown that adding discourse information on boots the performance of sentence –level paraphrase acquision, which consequently gives paraphrase fragments from matched sentence.

Hurgo Hernauilt and Helmut Predinger present HILDA discourse parser [1,2] based on RST and support `vector machine (SVM) classifier has been used for discourse segmentation and relation labeling. HILDA discourse parser [1] can parse entire text whereas publicly available parser SPADE (Soricut and Marcuu-2000) [10] is limited to sentence level analysis.

Stephen Soderland and Wendy Lehnert [11] also used discourse analysis for the information extraction. They described a system that learns discourse rules for domain specific analysis of unrestricted text. They involved complex series of decision about merging co-referential objects filtering and identifying logical relation between domain objects. Stephen Soderland and Wendy Lehnert [12] produce the Wrap-up algorithm for the information extraction which consist of two task discourse analysis which done at the text level. This discourse analysis mainly consists of co-reference resolution and 2nd task is information extraction from text using sentence analysis.

A Theory-Refinement [16] Approach to Information Extraction proposed by Tina Eliassi-Rad Jude Shavlik for information extraction. In Theory refinement partial domain knowledge is used for in formation extraction. This supervised learning may be incorrect. This approach uses generate and- test to address the IE task.

These all approaches use the statistic approach for the information extraction .This paper proposed linguistic approaches for information extraction. The linguistic method uses the relation between sentences for collecting the relevant data.

3. PROPOSED APPROACH

Basic approach of our proposed system is to use the discourse analysis for the information extraction. For system we are trying use the sentential relation proposed by RST theory. For the discourse analysis we have generate discourse parse tree implemented with RST theory. There are two building block processes for information extraction is one Discourse analysis and anotherone is information extraction. The system acchitechture of praposed system is given in Figure 1.

Discourse Analysis for generating relevant summary. This summary is generated on the basis of Nuclearity implemented in RST theory.

3.1 Discourse Analysis using Discourse Parser

The Discourse analysis has been done by generating discourse parse tree. The step for generating Discourse parse tree is given below:

- Text is segmented into EDU (Elementary Discourse Unit).
- Finding the Nucleus and satellite by RST theory in different EDU
- Hold two EDU using RST theory and modulate into the tree structure.
- Step 2 and step 3 are repeated until all span (EDU) is merged

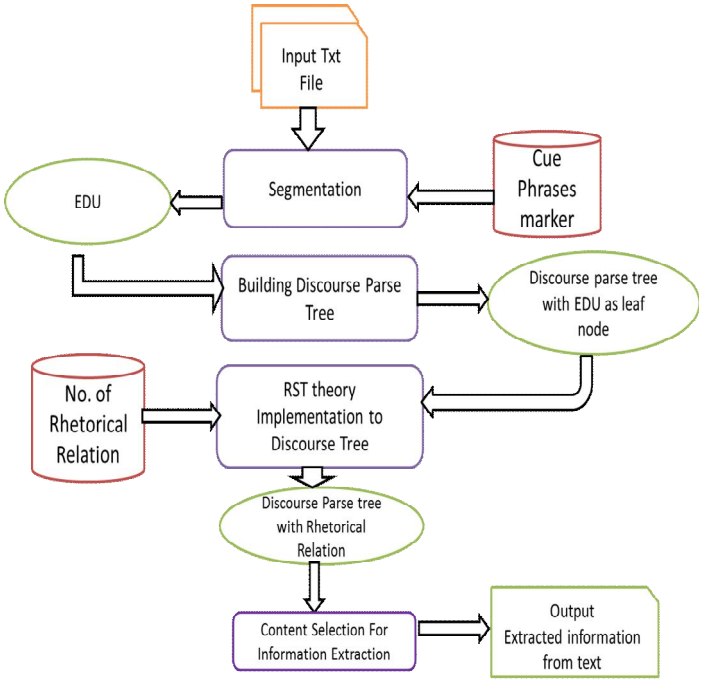

Figure 1: System Architecture

3.1.1 Segmentation

Elementary discourse unit (EDU) segmentation is an important process, since it separates full text into minimal discourse units that are used as an input of many applications such as text summarization, discourse parsing. Discourse segmentation is a part of discourse processing which separates full text into discourse units. The minimal discourse unit that was produced from discourse segmentation process is called elementary discourse unit (EDU) (Marcu, 1998, 1999). Many applications, such as text summarization (Marcu, 1999), discourse parsing (Polanyi et al., 2004) and machine translations, usually use EDUs as an input because sentences might be long discourse segments for these applications. EDU boundary has been defined on the basis of clause or clause like units. I have derive discourse structure of text using Rhetorical Structure Theory for relevant summary we will need to construct EDUs that need to determine rhetorical relation to understanding semantic of this text. If we select the most important units of text to be a summary, we select these units from EDUs-lists that are clause-like units. Therefore, correct and precise EDUs segmentation process is a significant process for relevant information.

For the EDU segmentation discourse cue marker and cue phrases has been used. I have used the number of key phrases in the discourse cue marker for making the segmentation.The oput of segmentation shown in figure 2.

```
<edu>BJP 's Prime Ministerial candidate Narendra Modi ,</edu>
<edu>under attack from critics for 2002 riots in Gujarat , on Saturday said</edu>
<edu>the country did not want " poison of communalism "</edu>
<edu>and needed safety and development instead . <s></edu>
<edu>Claiming that people were fed up of promises and empty talk ,</edu>
<edu>Modi ,</edu>
<edu>addressing a rally here ,</edu>
<edu>said</edu>
<edu>he had come with the " intent "</edu>
<edu>to deliver . <s></edu>
<edu>Targetting Congress ,</edu>
<edu>Modi said</edu>
<edu>it had made promises galore</edu>
<edu>but that alone could not bring development or change in the system . <s></edu>
<edu>" The country is fed up of promises ,</edu>
<edu>what matters is the intent .</edu>
<edu>I have come with the intent .</edu>
<edu>Congress has made scores of promises . <s></edu>
<edu>But people need development and not division , "</edu>
<edu>he told the gathering . <s></edu>
<edu>Trying to project himself as someone</edu>
<edu>who can usher in development ,</edu>
<edu>Modi said</edu>
<edu>people needed opportunities and not opportunism . <s></edu>
<edu>" People need skill</edu>
<edu>and not machinations .</edu>
<edu>They need jobs</edu>
<edu>and not politics .</edu>
<edu>They need safety and not the poison of communalism , "</edu>
<edu>he said . <s></edu>
<edu>Modi asked the gathering</edu>
```

Figure 2: Output of Segmentation

3.1.2 Building Discourse Parse Tree

For extracting the relevant information from the text discourse parsing has been done. In the discourse parsing EDU is used as input. According to the researcher all well-written text is supported by a hierarchically structured set of coherence relations which reflect the author intent. Thus these EDU organized hierarchically. Dependency parsing and other forms of syntactic analysis provide information on the grammatical structure of text at the sentential level. Discourse parsing, on the other hand, focuses on a higher-level view of text, allowing some flexibility in the choice of formal representation while providing a wide range of applications in both analytical and computational linguistics.

Thus we have implemented bottom-up construction of binary discourse parse tree. In discourse parse tree each EDU is represent the leaf node of the tree. But by using only discourse parse tree we will not get what will relevant or important content and what relation is present between content. For find the relevant information we are implemented RST theory on the discourse parser.Discourse parse tree for the input text is shown in figure 3

```
( Root (span 1 46)
  ( Nucleus (span 1 30)
    ( Nucleus (span 1 14)
      ( Nucleus (span 1 4)
        ( Nucleus (span 1 3)
          ( Nucleus (leaf 1) (text :BJP 's Prime Ministerial candidate Narendra Modi ,) )
          ( Nucleus (span 2 3)
            ( Satellite (leaf 2) (text :under attack from critics for 2002 riots in Gujarat , on Saturday said) )
            ( Nucleus (leaf 3) (text :the country did not want " poison of communalism ") )
          )
        )
        ( Nucleus (leaf 4) (text :and needed safety and development instead . <s>) )
      )
      ( Satellite (span 5 14)
        ( Nucleus (span 5 10)
          ( Satellite (leaf 5) (text :Claiming that people were fed up of promises and empty talk ,) )
          ( Nucleus (span 6 10)
            ( Nucleus (span 6 9)
              ( Satellite (span 6 8)
                ( Nucleus (span 6 7)
                  ( Nucleus (leaf 6) (text :Modi ,) )
                  ( Nucleus (leaf 7) (text :addressing a rally here ,) )
                )
                ( Nucleus (leaf 8) (text :said) )
              )
              ( Nucleus (leaf 9) (text :he had come with the " intent ") )
            )
            ( Satellite (leaf 10) (text :to deliver . <s>) )
          )
```

Figure 3: Discourse Parse Tree

3.1.3 Rhetorical Structure Theory Implementation

RST theory labeled EDUs with rhetorical relation. We have used the 23 relations for representing the rhetorical structure theory. These relations provide the concept of Nuclearity. The most frequent structural pattern is that two spans of text are related such that one of them has a specific role relative to the other. A paradigm case is a claim followed by evidence for the claim. We have used RST for analysis of the text. There is a graphical convention for expressing the structures of texts, but the particular claims made by the analyst can be made explicit based on the definitions of the relations and other structures of RST. The RST structure tree of input text shown in figure 4.

After the implementation of RST theory decision is available for Nucleus and satellite EDUs of text. The entire Nucleus has been extracted .The Nucleus is arranged according to their depth. By using human annotation gold standard it has found discourse Nucleus at superficial level provides more relevant information. The high depth Nucleus gives deep knowledge about text.

```
(textual-organization[N][N]
  (Explanation[N][S]
    (Explanation[N][S]
      (Joint[N][N]
        (same-unit[N][N]
          BJP 's Prime Ministerial candidate Narendra Modi ,
          (Attribution[S][N]
            under attack from critics for 2002 riots in Gujarat , on Saturday said
            the country did not want " poison of communalism "))
        and needed safety and development instead . <s>)
      (Explanation[N][S]
        (Contrast[S][N]
          Claiming that people were fed up of promises and empty talk ,
          (Enablement[N][S]
            (Attribution[S][N]
              (same-unit[N][N]
                (same-unit[N][N] Modi , addressing a rally here ,)
                said)
              he had come with the " intent ")
            to deliver . <s>))
        (same-unit[N][N]
          Targetting Congress ,
          (Contrast[N][N]
            (Attribution[S][N] Modi said it had made promises galore)
            but that alone could not bring development or change in the system . <s>))))
    (Contrast[S][N]
      (Attribution[N][S]
```

Figure 4: Rhetorical Structure Theory Tree

3.2 Information Extraction from discoursed Summary

After extracting the discourse based summary using the RST implemented discourse parse tree extraction module has been applied. In the extraction module supervised trained dictionary used as a bag of words.

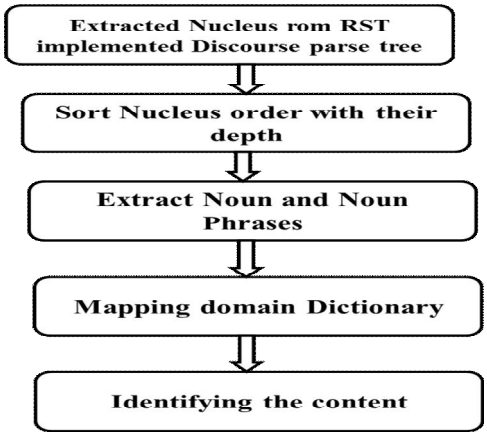

Figure 5: Steps of Information Extraction

What information present in newswires is generally depends on noun and noun phrases. These noun phrases are mapped into domain dictionary. The domain related dictionary has been used for modelling the topic of the text. The domain related dictionary has been play important role for modelling and sub-topic of text. Different steps of Information extraction is given in figure 5.

4. RESULT AND ANALYSIS

For the experiment purpose we have used the newswires of the BBC news of different domain. We have tested total number of 160 news which belongs to three different domains. We have considered three: Sports, Politics and Disaster. The domain is identified by using mapped word of text. Consider of newswires from BBC news.

Modi asked the gathering and needed safety and development instead . BJP 's Prime Ministerial candidate Narendra Modi , But people need development and not division , " Targetting Congress , to vote for BJP and its allies in Haryana " The country is fed up of promises , and shun other parties apparently weaving in a reference to ' spectacles ' but that alone could not bring development or change in the system . have reportedly made comments in the past the country did not want " poison of communalism " BJP has aligned with Kuldeep Bishnoi-led Haryana Janhit Congress in the state while leaders of Indian National Lok Dal he had come with the " intent " it had made promises galore people needed opportunities and not opportunism . " People need skill They need jobs and not machinations . said which is INLD 's election symbol . " It is only BJP 's poll symbol ' kamal ' Modi , They need safety and not the poison of communalism , " Trying to project himself as someone addressing a rally here

Figure 6: Extracted Politics Related Phrases

Example in figure 6 contain one of the news of BBC. The underlines word show that the news is belongs to category news of politics.

I have implemented experiment and finding out the result of for different domain classification. The strong bag of word domain is correctly classifying their domain. The domain of sports has their strong dictionary as compare to the other domain. Politics domain have some set of words which may not belong to these category.

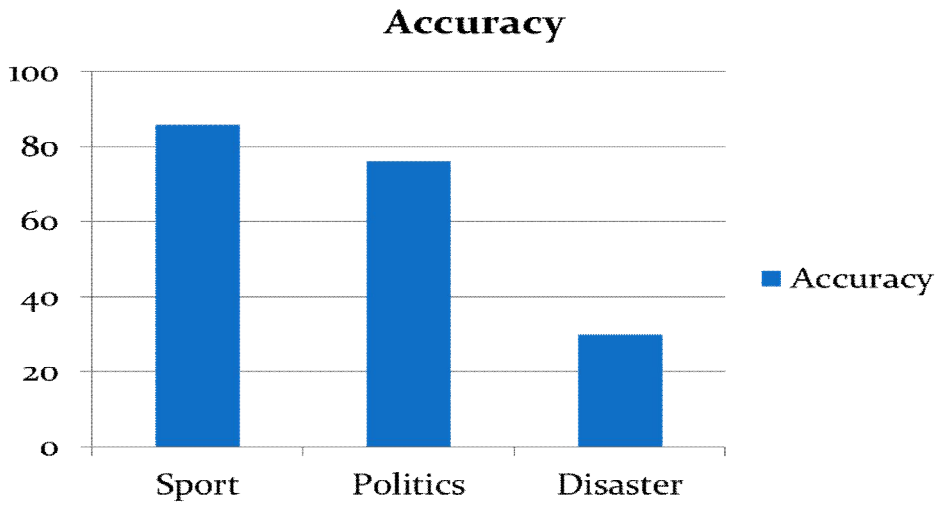

Figure 7: Result for Categarization of domain according to Discourse

We have implemented the above concept by using the discourse parsing .First segmentation has been done which is most important step for finding discourse content text. I have implemented segmentation process based on HILDA discourse parser [1] in which The discourse segmenter processes an input text one lexmes(word or punctuation mark) at a time and recognizes sentence and edu boundaries, and beginnings and ends of units. First, we measure the segmentation result when using parse trees from the Penn Treebank (Marcus et al., 1993) as our gold standard. Second, as a practical evaluation, we compare the performance when using parse trees generated respectively by the Stanford parser3 (Klein and Manning, 2003) and by the Charniak parser4 (Charniak, 2000).

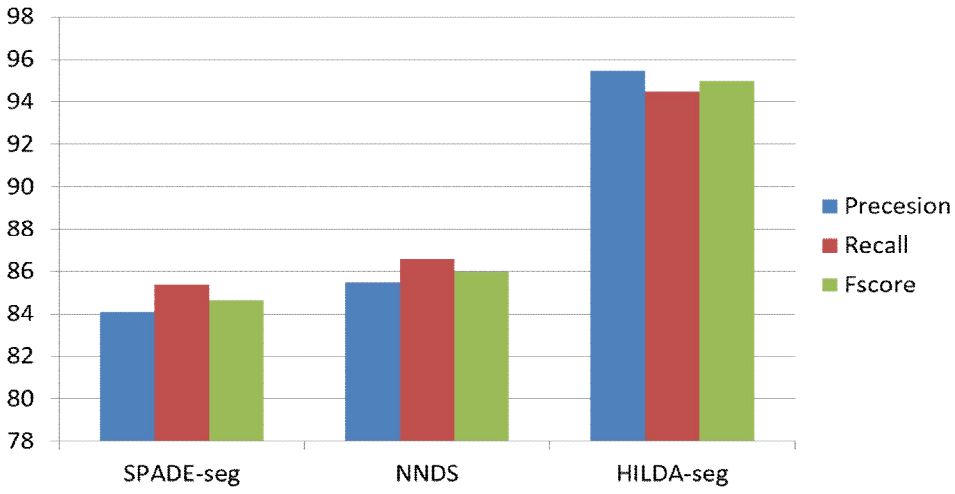

Figure 8: Analysis of Diiferent Segmentation Technique

5. CONCLUSION

We have experimented the information extraction using the discourse analysis.Discourse parsing can be used to remove the unwanted information from the the text. The short summary contian

the main relevant text for mextracting information. Bag of words of domain also play main role for content selection.

REFERENCES

[1] Helmut Prendinger, David A. duVerle, Mitsuru Ishizuka "HILDA: A Discourse Parser Using Support Vector Machine Classification "2/10; Accepted 11/10; Published online 12/10

[2] Vanessa Wei Feng,,Graeme Hirst "A Novel Discourse Parser Based on Support Vector Machine Classification" Conference on Computational Linguistics, pages 329–335

[3] Hammad Ali, Giuseppe Carenini, Gabriel Murray, and Raymond Ng " Designing a Discourse Parser for the Evaluative Text Genre" The Pacific Northwest Regional NLP Workshop (NW-NLP), 2010

[4] David A. duVerle, Helmut Prendinger "A Novel Discourse Parser Based on Support Vector Machine Classification" IJCNLP of the AFNLP, Suntec, Singapore, 2-7 August 2009

[5] Daniel Marcu "Discourse Tree Good Indicator for Important Text" S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[6] Daniel Marcu (1999). Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury editors, Advances in Automatic Text Summarization, pages 123-136, The MIT Press

[7] Rajen Subba, Barbara Di Eugenio "Automatic Discourse Segmentation using Neural Networks" Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 1-10, Cuiaba, MT, Brazil, October 24-26, 2011.

[8] Nynke van der Vliet, Gosse Bouma, Gisela Redeker"The automatic identification of discourse units in Dutch text" 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); Linköping Electronic Conference.

[9] Erick Galani Maziero and Thiago Alexandre Salgueiro Pardo "Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning" Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 1-10,Cuiaba, MT, Brazil, October 24-26, 2011

[10] Radu Soricut and Daniel Marcu (2003). "Sentence Level Discourse Parsing Using Syntatic and Lexical Information". Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), May 27-June 1, Edmonton, Canada.

[11] Stephen Soderland and Wendy Lehrent "Corpus-Driven Knowledge Acquisition for Discourse Analysis" Proceedings of the Twelfth National Conference on Articial Intelligence, 1994

[12] Stephen Soderland and Wendy Lehrent "Wrap-up: Trainable Discourse Module for Information Extraction" Journal of Artificial Intelligence and Research 2(1994) pg.131-158

[13] Stephen Soderland and Wendy Lehrent " Leaning Domain-Specific Discourse Rules for Information Extraction" AAAI 1995 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation

[14] MAITE TABOADA , WILLIAM C. MANN "Rhetorical Structure Theory: looking back and moving ahead" Discourse Studies SAGE Publications.(London, Thousand Oaks, CA and New Delhi) Vol 8(3): 423–459

[15] Helmut Prendinger , Paul Piwek , Mitsuru Ishizuka "Automatic Generation of Multi-Modal Dialogue from Text Based on Discourse Structure Analysis" International Conference on Semantic Computing

[16] Tina Eliassi-Rad,Jude Shavlik "A Theory-Refinement Approach to Information Extraction" Appears in the Proceedings of the 18th International Conference on Machine Learning (ICML 2001)

Authors

Ashwini Rahangdale: received Bachelor of Engineering Degree in Computer Science and Engineering from Maraathwada University Nanded India, and Master of Technology degree in Computer Science & Engineering from Shri Ramdeobaba College of Engineering & Management Nagpur, India in 2011 and 2014 respectively. Her research area is Natural Language Processing. She is the author of one research papers in International Conferences

Dr. A.J.Agrawal: Received Bachelor of Engineering Received Bachelor of Engineering Degree in Computer Technology from Nagpur University, India and Master of Technology degree in Computer Technology from National Institute of Technology, Raipur, India in 1998 and 2005 respectively. He received Ph.D. from Visvesvaraya National Institute of Technology, Nagpur, India in 2013. His research area is Natural Language Processing and Databases. He is having 15 years of teaching experience. Presently he is Assistant Professor in Shri Ramdeobaba College of Engineering & Management Nagpur, India He is the author of seven research papers in International and National Journal, Conferences.