Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

Discourse Analysis: Representation and Significance in Natural Language Understanding

Tiny Tonson Thomas
Department of computer engineering
K. J. Somaiya College of engineering
Mumbai, India

Pradnya S Gotmare
Department of computer engineering
K. J. Somaiya College of engineering
Mumbai, India

***Abstract*— Discourse analysis is one amongst the applications of Natural Language Processing. Discourse parsing is used for distinguishing the connectedness and specific talk relations among various units in a content. In this paper we are keen on Rhetorical Structure Theory(RST) among different theoretical frameworks of discourse parsing. It is semantically valuable strategy for depicting characteristic writings, portraying their construction in between parts of the content that hold relations with each other. This study is divided into two significant parts. In the initial segment we have examined about discourse segmentation and the subsequent part comprises of RST, rhetorical relations and an illustration of RST. The results demonstrates that discourse parsing will provide a proper solution for discourse analysis.**

Keywords*— RST relation, discourse analysis, syntactic tree, discourse segmentation

I. INTRODUCTION

Discourse analysis is a research method. It helps us to understand how a language can be used in our daily life situations. The word discourse means language in use. Discourse analysis is used for understanding written or spoken language and its relation to social context. This method is used in sociology, psychology, cultural studies etc.

Many analysts consider the larger discourse text so that we can perceive how it affects the entire means of the sentence. For example, consider a sentence like “People should use the toilet and not the pool, if there is not much inconvenience”. Now consider the second sentence like "Pool for members only.". Here we can see when we are taking the sentences independently we can see it as quite reasonable but when we are taking it altogether it makes us to switch the understanding of the main sentence after we have examined the second one.

Discourse will be either composed, as an example, in books, newspapers, streets, magazines, street signs or solicitations, or spoken, as an example, in discussions, verbal associations and television programs. Discourse Analysis (DA) is the logical structure which was made for examining real content and talk in the informative setting. A discourse is a bunch of implications through which a gathering of individuals impart about a specific subject. Discourse can be characterized in a thin or an expansive sense, in a narrow definition of discourse might refer only to communicate in or composed language.

II. LITERATURE SURVEY

In paper [1] a system for incorporating discourse data into machine perception applications is used. The model along acknowledges applicable sentences, sets up relations among them and predicts a solution. The model is evaluated employing a machine comprehension dataset consisting of fictional stories created by crowd workers. In this corpus, generally 50% of the questions rely upon various sentences within the section to form the proper answer. The authors have used a variety of lexical and syntactic features within the model. The StanfordCoreNLP package is used to pre-process the data. The data contains two specific sets: MC160 and MC500, which are of various size. Each section has four question, with four answer choices each. The questions are divided into two types: single, if the question can be answered utilizing a solitary sentence in the section, or multi otherwise. For every question, the system gains 1 point if it scores the highest correct answer and otherwise it scores 0. Just in case of ties, an inverse weighting scheme is used to assign partial credit. So, if there are three answers (including the proper one) it means tie for the best score and the system gains 1/3 points. The results demonstrated that MC160 test set, this model achieved best performance of 73.23% accuracy and MC500 test set model achieved 63.75% accuracy.

In paper [2] a sturdy answer reranking model for non-factoid answers is used. It comes with lexical semantics with discourse data that are: a shallow representation revolved with discourse markers and another is the deep one which is dependent on RST. The authors have used this system on two contents one from Answers within Yahoo and another within biology textbook data, and on 2 varieties of non-factoid questions which is: manner and reason. The proposed answer reranking segment is inserted in the Question-answering (QA) structure. Candidate Question-answering (CQA): In this situation, the undertaking is characterized as reranking all the user posted answers for a specific questions. Traditional Question-Answering: In this situation answers are powerfully built from bigger documents. The system is then assessed on multiple genres and question types and get benefits of up to 24% relative improvement over a solid benchmark that consolidates information retrieval and lexical semantics.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

The paper [3] is a study paper which uses the data of discourse structure (RST-DT). The structure of text is demonstrated in the form of tree in three different ways in RST framework. They are: (1) the leaves represents discourse units; (2) every node is shown by their own nuclearity; and (3) every node has a rhetorical relation within the texts. RST is a constituency-based theory: A global unit is formed by combining discourse units with discourse relations. The RST relation consists of a nucleus and satellite. This study utilizes Mean Discourse Distance equation to ascertain the mean for every RST text and then all the values sum up to mean discourse distance and divide this by RST texts.The authors have planned to elaborate this algorithm by obtaining data from RST-style corpus in various languages in the near future.

The paper [4] is a study paper which models the discourse construction of spontaneous spoken responses which is used for associate for non-local speakers for their English-speaking proficiency. This study has chosen data of 600 spoken responses, which were taken from TOEFL R Internet-based Test. The annotators portion a deciphered spoken relations into EDU spans of text. In this the authors have embraced an opensource parser. Results shows that the framework can reach an accuracy of 61.0%.

III. DISCOURSE SEGMENTATION

Discourse segmentation may be defined as determining the types of structures for large discourse. It is quite difficult to implement discourse segmentation, but it is very important for information retrieval, text summarization and information extraction kind of applications. Fig 3.1 shows the segmentation model in the form of a linear-chain Conditional Random field(CRF). Each and every sentence is shown in the form of a single linear chain.

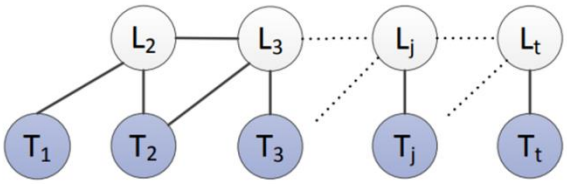

Fig 3.1: Segmentation model in the form of a linear CRF [14]

The layer 1 has many token nodes Ti's, $1 \leq i \leq t$, and then the layer 2 represents the label Li for every pair of tokens Ti−1 and Ti. For every neighbouring tokens in a sentence, i.e., Ti−1 and Ti, there is binary node Li which is used to represent the label of the pair, i.e., there exists a boundary in between the pairs: if we consider Li = B, an EDU boundary is inserted before Ti; and if we consider Li = C, then the two neighbouring tokens are taken as an EDU. Then a CRF-based model can be used for marking the entire arrangement of tokens in a sentence, in light of the fact that a CRF is fit for considering the consecutive data in the unique situation, and tackling the issue of deciding limits in a single pass.

IV. ROLE OF RHETORICAL STRUCTURE THEORY

Rhetorical Structure Theory (RST) is perhaps the most generally acknowledged frameworks for discourse analysis. RST is a useful hypothesis of text structure, which depicts rhetorical relations between parts of texts on various levels. RST is a framework created to represent text structure above the clause level. It shows how a text can be recursively broken into smaller portions called Elementary discourse units (EDUs).

Rhetorical relations in RST

Rhetorical relations are the central constructs in RST. Relation area units are characterized to hold two texts spans known as the nucleus and satellite which non-overlap with each other. One objective of this study is to create it easier so as to recognize the involvement of the analyst's judgement in the analysis.

RST provides a bunch of around 23 rhetorical relations, taxonomized on the informational and presentational premise. The rhetorical relations vary from one paper to different yet the central core of relations as introduced in Mann and Thompson (1987) given in the figure below:

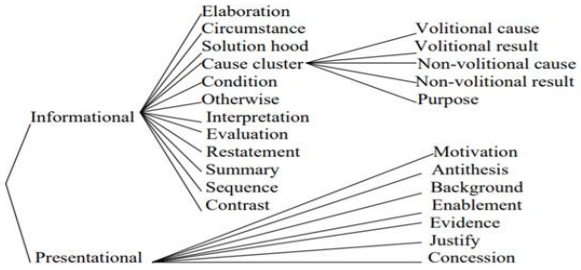

Fig: 4.1 Mann and Thompson's Rhetorical relations [19]

RST has five schemas as given in the diagram below:

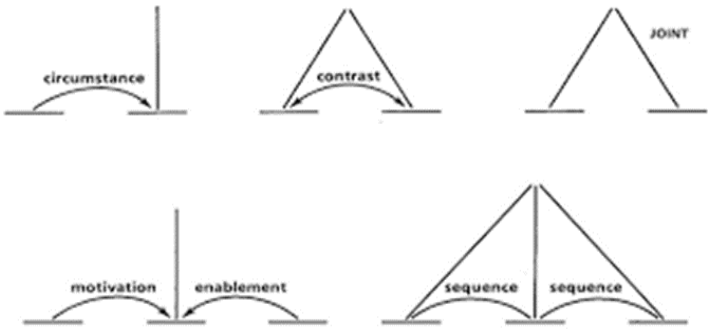

Fig: 4.2 Examples of five schemas [19]

An Example of RST-Style Discourse Tree Representation

The RST discourse Treebank is nothing but the data from Penn Treebank that is acquired from Wall Street Journal.

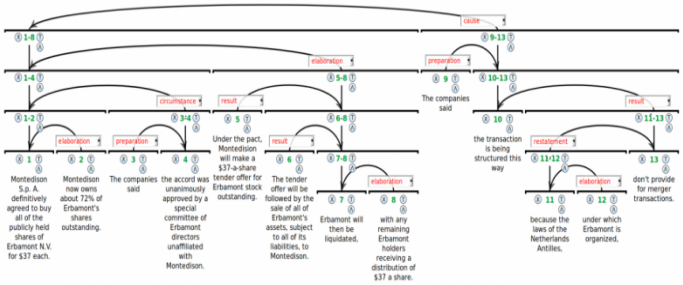

Fig: 4.3 RST tree (WSJ_0660) [3]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

In this the first one (1-2) is nucleus and another one is satellite. The leaves within the above are the text spans of a sentence, clause or EDU phrase, also known as elementary discourse units or EDUs in RST; these units are also known as discourse segments.

In the above example we can see every sentence is related to each other using rhetorical relations.

The node 1 has elaboration relation with node 2. Elaboration means the satellite gives additional information about the scenario given in the nucleus.

The node 3 has preparation relation with 4. Preparation means the satellite prepares the reader to expect and interpret the text given in the nucleus.

The node (1-2) and (3-4) has a circumstance relation with each other. The satellite in the circumstance relation sets a framework within which to interpret the nucleus.

The node 5 has result relation with node (6-8). The nucleus is an action carried out by an animate agent and the satellite is the reason for the nucleus.

The node (11-12) has a restatement relation with other node (13). Restatement means the satellite gives rexpression of the situation presented in the nucleus.

The node (1-8) has a cause relation with other node (9-13). Cause means the satellite gives another scenario which is created due to that one or with anyone's action which is mentioned in the nucleus.

The RST tree above is drawn to a dependency syntactic tree as given below:

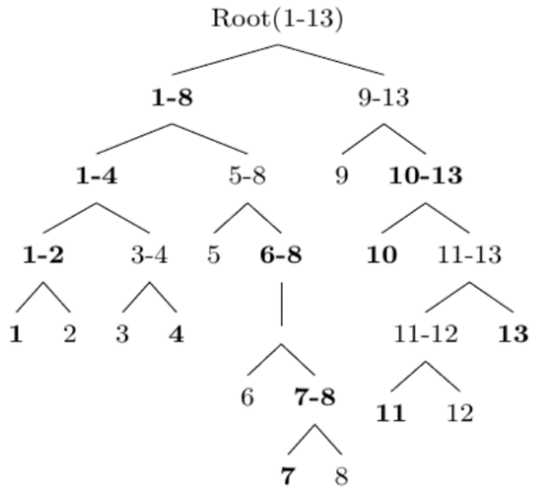

Fig: 4.4 Syntactic tree (WSJ_0660) [3]

Calculating discourse distance

For calculating discourse distance dependency distance algorithm is used. It is shown in the table below:

from(satellite)	to(nucleus)	frequency	relation
2	1	1	elaboration
3	4	1	preparation
(11-12)11	13	1	restatement
9	(10-13)10	1	preparation
(9-13)10	(1-8)1	1	cause

Table 4.1: Discourse distance [3]

Another Example of RST-Style Discourse Tree Representation

Below is a paragraph which has the examples of different rhetorical relations.

1. The hospital authorities had to control the rush
2. when many people rushed for Covid-19 vaccination at the hospital.
3. The hospital announced free vaccines for the poor and needy.
4. The rush in the line gives all of us a message, to get vaccinated and to protect yourself and everyone.
5. Every rule has exceptions,
6. but this scenario of people rushing up for vaccination demonstrates their awareness towards their own country and
7. not laziness.

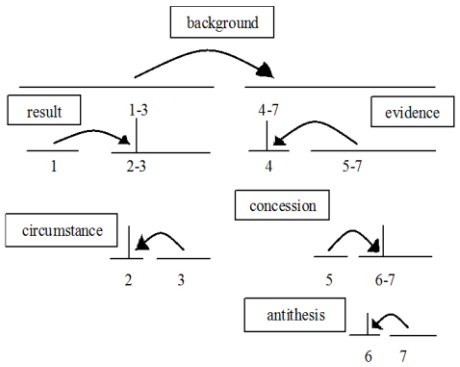

Fig 4.4 RST tree

Calculating discourse distance

from(satellite)	to(nucleus)	frequency	relation
3	2	1	circumstance
7	6	1	antithesis
1	(2-3)2	1	result
5	(6-7)6	1	concession
(5-7)6	4	1	evidence
(1-3)2	(4-7)4	1	background

Mean discourse distance = (3-2) + (7-6) + (2-1) + (6-5) + (6-4) + (4-2) / (7-1) = 1.33

V. CONCLUSIONS

The definitions within the paper provides an explicit and examinable understanding for a RST primary investigation. As a fascinating system, Rhetorical Structure Theory provides many features which is useful in discourse theories. It depicts

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

the relations between the text elements, recognizing the starting point of connection and the degree of things associated with it. RST is not affected by the size of text and hence it is applied to a wide range of text size. RST is semantically very helpful theory. The different relations used with RST tree like causal, elaboration, circumstance can be useful in machine comprehension applications, question-answering system etc and it is an engaging initial point for a wide range of studies.

REFERENCES

[1] Karthik Narasimhan, Regina Barzilay," Machine Comprehension with Discourse Relations", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing, IEEE, 2015.
[2] Mihai Surdeanu, Peter Jansen," Discourse Complements Lexical Semantics for Non-factoid Answer Reranking", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, IEEE, 2014.
[3] Kun Suna, Wenxin Xiongb," A Computational Model for Measuring Discourse Complexity", Association for Computational Linguistics, IEEE, 2019.
[4] Xinhao Wang, Binod Gyawali, James V. Bruno, Hillary R. Molloy, Keelan Evanini, Klaus Zechner," Discourse Modeling of Non-Native Spontaneous Speech Using The Rhetorical Structure Theory Framework", 2018 International Conference on Speech Communication, IEEE, 2018.
[5] J. Potter, D. Edwards," Discourse Analysis", Association for Computational Linguistics, IEEE, 2019.
[6] Hiroki Yamaguchi, Yukio Ohsawa, Yoko Nishihara," Discourse analysis for considering emotional effects on others", IEEE International Conference on Systems, Man and Cybernetics, IEEE, 2018.
[7] Eric N. Forsythand, Craig H. Martell, " Lexical and Discourse Analysis of online chat dialog", IEEE International conference on Semantic Computing", IEEE, 2018.
[8] Costin-Gabriel Chiru, Stefan Trausan-Matu," A tool for discourse analysis and visualization", IEEE International conference on emerging intelligent data and web technologies, IEEE, 2018.
[9] McCarthy, M. (2006) Discourse Analysis for Language Teachers. Cambridge: Cambridge University Press.
[10] https://shodhganga.inflibnet.ac.in/bitstream/10603/96162/12/12_chapter2.pdf
[11] https://www.oreilly.com/library/view/natural-language-processing/9781787285101/ch28.html
[12] https://shodhganga.inflibnet.ac.in/bitstream/10603/96162/11/11_chapter1.pdf
[13] https://corpling.uis.georgetown.edu/rstweb/info/
[14] http://ftp.cs.toronto.edu/pub/gh/Feng-thesis-2015.pdf.
[15] https://oiipdf.com/mastering-natural-language-processing-with-python
[16] https://www.scribbr.com/methodology/discourse-analysis/
[17] https://www.emeraldgrouppublishing.com/how-to/research/data-analysis/use-discourse-analysis
[18] https://www.thoughtco.com/discourse-analysis-or-da-1690462
[19] https://www.sfu.ca/rst/05bibliographies/bibs/Mann_Thompson_1988.pdf
[20] https://www.birmingham.ac.uk/Documents/college-artslaw/cels/essays/appliedlinguistics/RepresentationsofMeaning-DissertationMPost.pdf