Building Discourse Parser for Thirukkural

R. Anita, C.N. Subalalitha
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Kattankulathur-603 203, Tamilnadu, India
{anitaramalingam17,subalalitha}@gmail.com

Abstract

Thirukkural is one of the famous Tamil Literatures in the world. It was written by Thiruvalluvar, and focuses on ethics and morality. It provides all possible solutions to lead a successful and a peaceful life fitting any generation. It has been translated into 82 global languages, which necessitate the access of Thirukkural in any language on the World Wide Web (WWW) and processing the Thirukkural computationally. This paper aims at constructing the Thirukkural Discourse Parser which finds the semantic relations in the Thirukkurals which can extract the hidden meaning in it and help in utilizing the same in various Natural Language Processing (NLP) applications, such as, Summary Generation Systems, Information Retrieval (IR) Systems and Question Answering (QA) Systems. Rhetorical Structure Theory (RST) is one of the discourse theories, which is used in NLP to find the coherence between texts. This paper finds the relation within the Thriukkurals and the discourse structure is created using the Thirukkural Discourse Parser. The resultant discourse structure of Thirukkural can be indexed and further be used by Summary Generation Systems, IR Systems and QA Systems. This facilitates the end user to access Thirukkural on WWW and get benefited. This Thirukkural Discourse Parser has been tested with all 1330 Thirukurals using precision and recall.

1 Introduction

Tamil literature has so many nuggets hidden in it which need to be explored for the goodness of the society. One of the ways to explore the Tamil literature is to make it easily accessible on the World Wide Web (WWW). For instance, Thirukkural is one of the famous Tamil literatures in the world and it is respected by people across the globe. In order to make it to reach to all people, it should be made available on the web. This makes necessary to process it computationally. Hence, this paper proposes a methodology to perform a discourse analysis of Thirukkural, which aids in exploring its semantics and also organizing it on the web.

Natural Language Processing (NLP) is the process of interaction between computer and human or natural languages. Analysis of text can be done at various levels namely, word, clause, sentence, paragraph and document. Discourse analysis is used for analyzing the text beyond the clause level. The proposed work attempts to extract the relations found within the Thirukkural.

Discourse structure of a text can be built by using a popular theory called, Rhetorical Structure Theory (RST) (Thompson and Mann, 1987; Mann and Thompson, 1988). Using RST-based discourse relations, the RST captures the coherence among the Natural Language (NL) text spans. The coherence can be found between two or more text fragments. The text fragments could be within a sentence, across sentences, across paragraphs and even across documents.

In this paper, each Thirukkural is considered as a sentence and discourse parser is built using RST. The contributions of this paper are twofold.

1) Finding feature set using rule based approach.
2) Building Discourse parser by identifying discourse relations.

D M Sharma, P Bhattacharyya and R Sangal. Proc. of the 16th Intl. Conference on Natural Language Processing, pages 18–25
Hyderabad, India, December 2019.

The rest of the paper is organized as follows. Section 2 describes background details. Section 3 describes the related work. Section 4 discusses the proposed work. Section 5 gives details on the evaluation of the proposed technique. Section 6 presents the conclusion and future works.

2 Background

This section describes about the Thirukkural and the basics of RST based Discourse Parsing.

2.1. Thirukkural

Thirukkural consists of 1330 couplets or Thirukkurals. They are classified into three sections and 133 chapters. Each chapter in Thirukkural has a specific subject and consists of ten couplets or Thirukkurals. A couplet consists of two lines. Each Thirukkural or couplet is formed with seven cirs (words). First line of the couplet consists of four cirs and the second line of the couplet consists of three cirs. A single Tamil word or a combination of two or more Tamil words forms a cir.

2.2. Rhetorical Structure Theory

RST is a descriptive theory which focuses on the organization of the natural language. It was proposed by Bill Mann, Sandy Thompson, and Christian Matthiessen at the University of Southern California (Thompson and Mann, 1987; Mann and Thompson, 1988). It identifies the coherence between the text spans using discourse relations and forms a discourse structure called rhetorical structure. The discourse units are Nucleus, Satellite and Discourse Relations. The nucleus carries the necessary information and the satellite carries the additional information supporting the nucleus.

Discourse relations are organized into three categories, namely, subject matter, presentational, and multinuclear. In subject matter relations, satellite is a request or problem posed by the reader, i.e. satisfied or solved by nucleus. Elaboration, evaluation and condition are some of the subject matter relations. In presentational relations, satellite increases reader's inclination in accepting the facts stated in nucleus. Antithesis, background and enablement are some of the presentational relations. In multinuclear relations, two nuclei are connected instead of one nucleus and one satellite. Conjunction, contrast and sequence are some of the multinuclear relations.

Figure 1 shows an example of how the nucleus, satellite, and the discourse relation are identified for an English sentence in Example 1.

Example 1 Raj sings well but he could not win the contest.

Nucleus:	Raj sings well
Satellite:	he could not win the contest
Discourse Relation:	Antithesis

Figure 1. NRS Sequence for Example 1.

In Example 1, the sentence holds antithesis relation. It is identified by the signal word *but*. "Raj sings well" is the nucleus, because it represents the ideas favored by the author. "He could not win the contest" is the satellite, because it represents the ideas disfavored by the author. These NRS sequences capture the inherent semantics in the texts which is applied to the Thirukkural couplets by the proposed approach.

3 Related Work

Subba and Di (2009) found discourse relations by using shift reduce parsing model and WordNet. The linguistic cues were used as features. The document was analyzed at sentence level. Hernault et al. (2010) constructed discourse parser by building discourse tree using Support Vector Machine Classifier. The document was analyzed beyond the sentence level and the combination of syntactic and lexical features such as words, POS tags and lexical heads were used as feature sets.

Hernault et al. (2010a) used a semi-supervised method called Feature Vector Extension for discourse relation classification. The method was based on the analysis of co-occurring features present in unlabelled data, which was then taken into account for extending the feature vectors given to a classifier. The word pairs, production rules from parse trees and Lexico-Syntactic context at the border between two units of text were used as features for the algorithm.

Sucheta et al. (2011) identified explicit discourse connectives for Penn Discourse Tree Bank (PDTB). They proposed shallow discourse parsing for performing token level argument segmentation. The document was analyzed at sentence level. The lexical, syntactic and semantic features were used as features.

Sucheta et al. (2012) improved a shallow discourse parser by using a constraint-based method based on conditional random fields and the recall was improved. Sucheta, Giuseppe, and Richard (2012) constructed a parser which uses local constraints and then global constraints. They analyzed the text at the inter sentence level and they used the lexico-syntactic features.

Subalalitha and Ranjani Parthasarathi (2013) used Tamil and Sanskrit literature concepts called suthras and sangatis, along with the current-day text processing theories namely, RST, Universal Networking Language for identifying semantic indices for Tamil documents. Suthras are used for representing the text in a crisp manner.

Sobha et al. (2014) and Sobha and Patnaik (2004) proposed automatic identification of connectives and their arguments for the Indian languages Hindi, Malayalam and Tamil. They used Conditional Random Fields machine learning technique. They used 3000 sentences from a health domain as a corpus. Sobha et al. (2014), annotated the three language corpus, namely Tamil Hindi and Malayalam, with the discourse relations.

Lin et al. (2014) constructed an end-to-end discourse parser in the PDTB style. Their parser identified all discourse and non-discourse relations, labeled the arguments, and found the sense of relation between arguments. The document was analyzed at paragraph level. The lexical, syntactic and semantic features were used as features.

Yangfeng and Jacob (2014) transformed surface features into a latent space by using a representation learning approach that facilitates RST discourse parsing. They used shift reduced discourse parser and analyzed the document at sentence level.

Uladzimir et al. (2015) segmented the German text for the RST-based discourse parsing. They analyzed the text at sentence level. Parminder et al. (2015) proposed document level sentiment analysis using RST discourse parsing and recursive neural network. They analyzed the text at document level and lexical features were used as features.

Subalalitha and Ranjani Parthasarathi (2015) found 13 RST Relations in Tamil documents. The Naïve Bayes probabilistic classifier machine learning algorithm was used and the Tamil documents were analyzed beyond the sentence level. The high level semantic features were used by their discourse parser, which were inherited from UNL to construct rhetorical structure trees.

Manfred et al. (2016) annotated the corpus with two theories, namely, RST and Segmented Discourse Representation Theory. It was also annotated with the argumentation annotation. The document was analyzed at sentence level. The syntactic and semantic features were used as feature set.

Yangfeng et al. (2016) proposed a latent variable recurrent neural network for finding the discourse relation between adjacent sentences. They analyzed the text at inter sentence level and they have used lexical features. Yangfeng and Noah (2017) proposed text categorization by using recursive neural network and RST. The document was analyzed at sentence level.

It can be observed that, the existing discourse methodologies analyzed the text in English documents and expository type Tamil documents. This paper proposes a discourse methodology that makes use of RST to identify the semantic relations/discourse relations from a Tamil literature text which lacks a regular pattern for semantic analysis. Unlike English which has a fixed SVO (Subject Verb Object) sentence pattern, Tamil expository texts have either SVO or SOV (Subject Object Verb) pattern. Tamil literatures on the other side neither follow SOV nor SVO pattern. Tamil literatures also have a relatively rich set of morphological variants (Anand et al., 2010; Goldsmith, 2001). This makes the processing of Tamil literature more complex than processing the expository Tamil documents. This paper focuses on finding discourse relations in a Tamil literary work called, Thirukkural, which has the structure of classic Tamil language poetry form, called venba. Venba style Tamil literature consists of lines between two and twelve. Expository Tamil documents have the cue words in middle of the sentence. It is not difficult to find the nucleus satellite identification for expository type of texts, whereas the cue words in Tamil literature specifically in venba style of texts will be present in any part of the sentence. If the cue word is present in the middle of the Thirukkural couplet, it is not difficult to find the nucleus satellite identification. If it is present at either end of the Thirukkural couplet then it is difficult to find the nucleus satellite identification. The proposed

Thirukkural Discourse Parser handles these cases to an extent which is discussed in the upcoming sections.

4 Proposed Work

The Tamil Thirukkural couplets are given as input. Initially the cue phrases or signal words are identified in each Thirukkural. RST based discourse relations are identified by Thirukkural Discourse Parser based on the cue words and semantics. Then the Nucleus and Satellites are identified for each Thirukkural. Finally, the NRS sequences are identified as output from the Thirukkural Discourse Parser.

4.1. Feature Sets

The connectives connecting two clauses of the Thirukkural are used as the feature set as they signal a discourse relation. An analysis of the 600 Thirukkural has been done and the feature set for each discourse relation has been identified. For condition relation, 108 features have been identified; for evidence relation, 36 features have been identified; for contrast relation, 37 features have been identified; for enablement relation, 24 features have been identified; for background relation, 9 features have been identified. The feature sets, cue words and signal words are interchangeably used in this paper. The part of the cue words are appeared in Table 1. For example, ‘இலவே (Ilave-If not)’, ‘ஆயின் (Ayin-If)’, and ‘ஆற்றின் (Arrin-If someone did)’ are some of the cue words commonly appeared in Thirukkural.

4.2. Discourse Relation Identification

The discourse relations namely, condition, evidence, contrast, enablement and background are identified by the Thirukkural Discourse Parser. A cue word may either be a single word which can be explicitly identified by the Thirukkural Discourse Parser or it may be a case suffix which may have to be split by the morphological analyzer (Anandan et al., 2001).

If the cue words explicitly appear in the Thirukkural, then the RST based discourse relations are identified using the signal words in Table 1. The cue words are given in Tamil along with their English transliteration and English meaning.

If the cue words do not explicitly appear in the Thirukkural, then the morphological analyzer is used for finding the cue words. For example, in the word ‘எழுத்தெல்லாம் (Eluttellam-All the letters) ’, the cue word ‘எல்லாம் (Ellam-Everything)’ is a case suffix and so the morphological analyzer is used to isolate the cue word (‘எழுத்து+ எல்லாம்’). Now the cue word ‘எல்லாம் (Ellam- Everything)’ can be used for identifying the RST based discourse relation.

S. No.	Relation	List of Cue words
1	Condition	இலவே (Ilave-If not), என்னும் (Ennum- The), பெறின் (Perin- If received), என்னாம் (Ennam-If), ஆயின் (Ayin-If), ஆற்றின் (Arrin-If someone did)
2	Evidence	வேண்டா (Venta- Do Not), அதுவல்லது (Atuvallatu - That is not), தான்(Tan- Just), போன்று (Ponru- Like), தேரின் (Terin- Selection), எங்ஙனம் (Ennanam- How)
3	Contrast	அரிய (Ariya – Rare), மற்றெல்லாம் (Marrellam- On every), ஆதல் (Atal- Therefore), உய்க்கும் (Uykkum- That derived)
4	Enablement	எல்லாம் (Ellam- Everything), அவருள்ளும் (Avarullum- Plunge), மன்ற (Manra- House), போல(Pola-Like)
5	Background	எனினும் (Eninum- However), செல்லாது (Cellatu- Invalid), இனிதே (Inite- Greeter), அற்று (Arru- Without), உறையும் (Uraiyum- Freezing)

Table 1. Some Relations and Cue Words

4.3. Nuclearity Identification

The cue words are appeared anywhere in the Thirukkural. In most of the Thirukkural, it is appeared in the middle. The text before the cue word is considered as Clause1 and the text after the cue word is considered as Clause2. Clause1 and Clause2 can be indicated as nucleus and satellite.

In few Thirukkural, Clause1 can act as the nucleus and Clause2 can act as the satellite. And in others, they may be vice versa. Hence, it is identified separately and the NRS sequences are identified accordingly. Table 2 lists some of the cue words appear in Thirukkural in which Clause1 and Clause2 are categorized as nucleus and satellite.

Example 2 in Figure 2 shows NRS sequence identified by the Thirukkural Discourse Parser.

S. No	Relation	Clause1-Nucleus, Clause2-Satellite	Clause2-Nucleus, Clause1-Satellite
1	Condition	இலவே (Ilave-If not), என்னும் (Ennum-The), பெறின் (Perin- If received), என்னாம் (Ennam-If), வைப்பின் (Vaippin-Fund)	ஆயின் (Ayin-If), ஆற்றின் (Arrin-If someone did)
2	Evidence	வேண்டா (Veṇṭa- Do Not), அதுவல்லது (Atuvallatu - That is not),	தான் (Tan- Just), போன்று (Ponru- Like), தேரின் (Terin- Selection), எங்ஙனம் (Ennanam- How)
3	Contrast	அரிய (Ariya – Rare), மற்றெல்லாம் (Marrellam- On every), ஆதல் (Atal- Therefore), உய்க்கும் (Uykkum- That derived)	உரியர் (Uriyar- Belong) , செயினும் (Ceyinum- Though did)
4	Enablement	எல்லாம் (Ellam- Everything), அவருள்ளும் (Avarullum- Plunge), மன்ற (Manra- House)	போல (Pola-Like)
5	Background	செல்லாது (Cellatu- Invalid), இனிதே (Inite-Greeter), அற்று (Arru- Without), உறையும் (Uraiyum- Freezing)	எனினும் (Eninum- However),

Table 2. Nucleus and Satellite Identification

The Thirukkural in Figure 2 is given as the input for the Thirukkural Discourse Parser. This Thirukkural has the cue word 'போல (Pola-Like)' explicitly. This cue word is used for identifying the Enablement relation. The Thirukkural Discourse Parser identifies 'Enablement' as the discourse relation, 'ஆங்கே இடுக்கண் களைவதாம் நட்பு' as nucleus, (as it contains an action), and 'உடுக்கை இழந்தவன் கை' as satellite, (as it contains the information for performing the action). Similarly NRS sequences for all such cases present in the Thirukkurals are identified by using the Thirukkural Discourse Parser. Figure 3 shows Nucleus, Satellite and Discourse relation for the Example 2.

Example 2:
உடுக்கை இழந்தவன் கைபோல ஆங்கே இடுக்கண் களைவதாம் நட்பு.

English Transliteration:
Utukkai ilantavan kaipola anke itukkan kalaivatam natpu.

Meaning in English:
True friendship hastens to the rescue of the afflicted as readily as the hand of one whose garment is loosened.

Figure 2. Example 2

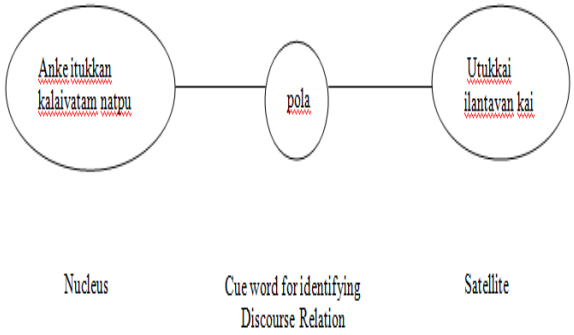

Figure 3. NRS Sequence for Example 2.

The algorithm for Thirukkural Discourse Parser is shown in Figure 4.

Input: Thirukkural Couplets
Output: NRS Sequences
(i) Find cue words in all Thirukkural Couplets
(ii) Store it separately corresponding to the relation
(iii) **for** Thirukkural couplets = 1 to 1330 **do**
if cue word is present in the Thirukkural **then**

identify the Discourse relation
display the NRS sequences
end
if cue word is not present in the Thirukkural **then**
use morphological analyzer to find the cue word

identify the Discourse relation
display the NRS sequences
end
end

Figure 4. Algorithm for Thirukkural Discourse Parser

5 Evaluation

The features are extracted from 600 couplets. The 1330 Thirukkural couplets are given as the input for the Thirukkural Discourse Parser. The NRS sequences are identified by the Thirukkural Discourse Parser based on the cue words. This work is evaluated using the parameters, precision and recall. Precision (P), and recall (R) values are calculated using equations (1) and (2).

$$\text{Precision(P)} = \frac{\text{Number of relevant NRS sequences retrieved, (C)}}{\text{Total number of NRS sequences retrieved,(M)}} \quad (1)$$

$$\text{Recall(R)} = \frac{\text{Number of relevant NRS sequences retrieved, (C)}}{\text{Total number of relevant NRS sequences present, (N)}} \quad (2)$$

The Table 3 shows the precision and recall of the discourse parser. The total number of NRS sequences retrieved by the proposed Thirukkural Discourse Parser is denoted as - M, total number of relevant NRS sequences actually present in Thirukkural is denoted as - N, and number of relevant NRS sequences retrieved by the proposed Thirukkural Discourse Parser is denoted as - C. The value of the variables, C and N are calculated using human judgement. About six domain experts have calculated these metric N, and the average has been taken and presented in Table 3.

Relation	**N**	**M**	**C**	**P=$\frac{C}{M}$ (%)**	**R=$\frac{C}{N}$ (%)**
Condition	547	616	514	83.44	93.97
Evidence	248	283	225	79.51	90.73
Contrast	262	216	182	84.26	75.21
Enablement	189	158	129	81.65	76.33
Background	176	145	114	78.62	73.08

Table 3. Precision, Recall and F-Measure for the Discourse Relation

It can be observed from the table that the total number of NRS sequences relevant to the condition relation is 547 and total number of NRS sequences retrieved from the Thirukkural Discourse Parser is 616. This is because more than one cue words belonging to the Condition relation are appeared in some Thirukkurals.

In Example 3 shown in Figure 5, two cue words "*ஆற்றின்* (Arrin-If someone did)" and "*பெறின்* (Perin- If received)" have appeared in the Thirukkural. Both cue words belong to the Condition relation. In order to analyze the significance of each feature, the NRS sequences emerging out of two features present in the same Thirukkural is counted and hence, M is greater than N. Similarly, for Evidence relation also M is greater than N. On the other side, in Contrast, Enablement and Background relations, N is more than M.

Example 3:
செறிவறிந்து சீர்மை பயக்கும் அறிவறிந்து ஆற்றின் அடங்கப் பெறின்.

English Transliteration:
Cerivarintu cirmai payakkum arivarintu aarrin atankap perin.

Meaning in English:
Knowing that self-control is knowledge, if a man should control himself, in the prescribed course, such self-control will bring him distinction among the wise.

Figure 5. Example 3.

The correctly retrieved Thirukkurals, C is smaller than N in all the discourse relation, this is due to the inability of the discourse parser to extract the NRS sequences using implicit cue words.

Example 4:
உள்ளற்க உள்ளம் சிறுகுவ கொள்ளற்க அல்லற்கண் ஆற்றறுப்பார் நட்பு.

English Transliteration:
Ullarka ullam cirukuva kollarka allarkaṇ arraruppar natpu.

Meaning in English:
Do not think of things that discourage your mind, nor contract friendship with those who would forsake you in adversity.

Figure 6. Example 4.

In Example 4 shown in Figure 6, the cue word "போல (Pola-Like)" is implicit. The Enablement relation identification needs additional semantic analysis which is currently not done by the discourse parser.

In some Thirukkurals, more than one cue words pointing to different relations have appeared. Therefore more than one NRS sequences are identified by the Thirukkural Discourse Parser for the same Thirukkural.

In Example 5 shown in Figure 7, two cue words namely, "ஆற்றின் (Arrin-If someone did)" and "போல் (Pol-Like)" are present.

"ஆற்றின் (Arrin-If someone did)" is a cue word related to Condition relation and "போல் (Pol-Like)" is a cue word related to Evidence relation. So the Condition and Evidence relations are identified by the Thirukkural Discourse Parser.

Example 5:
ஒருமையுள் ஆமைபோல் ஐந்தடக்கல் ஆற்றின்
எழுநம்யும் ஏமாப் புடைத்து.

English Transliteration:
Orumaiyul amaipol aintatakkal arrin
elunamyum emap putaittu.

Meaning in English:
Should one throughout a single birth, like a tortoise keep in his five senses, the fruit of it will prove a safe-guard to him throughout the seven-fold births.

Figure 7. Example 5.

The precision and recall values of the discourse parser can further be increased by increasing the feature sets, by incorporating a machine learning algorithm. The efficiency can be increased by finding the discourse relations for the Thirukkurals having implicit cue words. The efficiency of the Thirukkural Discourse Parser also depends on the efficiency of the morphological analyzer. A high level semantic knowledge base such as WordNet (George, 1995) or ontology may improve the efficiency even better.

6 Conclusion and Future Works

Thirukkural has much valuable information that is to be followed by the society. In order to access the Thirukkural on the web, the semantic analysis of the same becomes necessary. This paper makes use of discourse theory, named, RST to do a discourse/semantic analysis on the Thirukkural which will be useful to retrieve the Thirukkural using an Information Retrieval System.

Keyword-based Thirukkural search is available but limits the user to retreive the Thirukkural containing only the query words. In this paper, we propose a methodology to construct a discourse parser for Thirukkural which aids in semantic analysis and semantic representation of Thirukkural. This kind of semantic representation can be used for efficient semantic indexing of Thirukkural for better retrieval.

Using the results of the proposed discourse parser, an analysis of how Thirukkural is written and organized is also evident which can be useful to write similar works in future. This kind of analysis can also help in automatic author detection of text (Stamatatos, 2008).

This paper focuses on the discourse relations present within a Thirukkural couplet. It may be further extended to find the relations across the Thirukkural couplets. An application that clearly depicts the discourse structure representation is also to be done.

References

Anandan, P., Ranjani Parthasarathy, and Geetha, T.V. 2001. Morphological Analyzer for Tamil. *ICON 2002*, RCILTS-Tamil, Anna University, India.

Anand Kumar, M., Dhanalakshmi, V., Soman, K.P., and Rajendran, S. 2010. A Sequence Labeling Approach to Morphological Analyzer for Tamil Language. *International Journal on Computer Science and Engineering (IJCSE)*, Vol. 02, No. 06, 1944-195.

George, A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11:39-41.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27, 153–198.

Hernault, H., Bollegala, D., and Ishizuka, M. 2010a. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics*: Cambridge, MA, pp. 399-409.

Hernault, H., Predinger, H., Duverle, D.A., and Ishizuka, M. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, vol.1, no.3, pp. 1-33.

Lin, Z., NG, H., and Kan, M. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2), 151-184.

Manfred Stede, Stergos Afantenos, Andreas Peldszus,Nicholas Asher, and Jeremy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portoroz.

Mann, W.C. and Thompson, S.A., 1988. Rhetorical structure theory: Toward a functional theory of

text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3), pp.243-281.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *EMNLP*, pages 2212–2218, 2015.

Polanyi, L.C., Culy, M.H., Van Den Berg, Thione, G.L., and Ahn, D. 2004. Sentential structure and discourse parsing. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain, pp. 80-87.

Sobha Lalitha Devi and B. N. Patnaik (2004), Discourse Connectives and Their Arguments in Malayalam", *24th South Asian Language Analysis*, University of Stony Brook , New York, November 19-21

Sobha Lalitha Devi, Lakshmi S and Sindhuja Gopalan(2014). "Discourse Tagging for Indian Languages", In *A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing*, Springer LNCS Vol 8403, pp. 469-480.

Sobha Lalitha Devi, Sindhuja Gopalan, Lakshmi S (2014). "Automatic Identification of Discourse Relations in Indian Languages", In *proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation*, Organized under LREC2014, Reykjavik, Iceland

Stamatatos, E. 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, 44(2), 790–799.

Subalalitha, C.N., and Ranjani Parthasarathi 2013. A Unique Indexing Technique for Indexing Discourse Structures. *Journal of Intelligent Systems*. Volume 23, Issue 3, Pages 231–243.

Subalalitha, C.N., and Ranjani Parthasarathi. 2015. Building a Language-Independent Discourse Parser using Universal Networking Language. *Computational Intelligence*, 31: 593–618.

Subba, R., and Di Eugenio, B. 2009. An effective discourse parser that uses rich linguistic information. In *The Annual Conference of the North American Chapter of the ACL*, Boulder, CO, pp. 566-574.

Sucheta Ghosh, Giuseppe Riccardi and Richard Johansson 2012 Global features for shallow discourse parsing. In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 150–159

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi and Sara Tonelli 2011 Shallow discourse parsing with conditional random fields. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pp. 1071–1079, Chiang Mai, Thailand, November

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi and Sara Tonelli 2012 Improving the recall of a discourse parser by constraint-based postprocessing. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. 2015. Discourse Segmentation of German Texts. *Journal of Language Technology and Computational Linguistics (JLCL)*, 30(1):71–98.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Yangfeng Ji, and Noah A. Smith. (2017). Neural discourse structure for text categorization. *arXiv preprint* arXiv:1702.01829.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *NAACL-HLT*.