An Approach to Discourse Parsing using sangati and Rhetorical Structure Theory

Subalalitha C N, Ranjani Parthasarathi
Dept of IST,
Anna University, Chennai – 600 025
subalalitha@gmail.com, rp@annauniv.edu

ABSTRACT

Sanskrit literature has many nuggets that could be applied to modern linguistic applications. One such nugget is the concept of sangati. Sangati expresses continuity and proper positioning of piece of text which is similar to the modern Rhetorical Structure Theory (RST). We propose two discourse parsers namely sangati based discourse parser and RST-Sangati based discourse parser. The proposed discourse parsers are extensions of the existing RST based discourse parser. We have used Naive Bayes probabilistic classifier for discourse relation and sangati labelling. We have tested our discourse parsers using 500 Tamil tourism domain specific documents and 21 RST- Discourse Tree (RST-DT) English documents. We have compared the performance of both the proposed discourse parsers and observed that when RST and sangati are used in union, the performance of the discourse parser is better. Also, we have done a performance comparison with two existing discourse parsers and have shown better performance.

Keywords: sangati, Discourse parser, Rhetorical Structure Theory, Universal Networking Language

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 73–82,
COLING 2012, Mumbai, December 2012.

1. Introduction

With the massive increase in documents in World Wide Web (WWW), the knowledge present in those documents needs to be managed properly. Rhetorical Structure Theory (RST) is a well known text representation technique that represents the knowledge present in the text using semantic relations known as discourse relations (Mann et al, 1988). RST captures the coherence between the text using the discourse relations. Similar to RST, in Sanskrit literature, a concept known as sangati exists which expresses continuity between texts. Using sangati and RST, in this paper, we propose two discourse parsers namely, sangati based discourse parser and RST-Sangati based discourse parser. Both the discourse parsers are extensions of our previous work on RST based discourse parser (Subalalitha et al, 2011). Consequently, we have compared RST and sangati and have observed that RST and sangati complement each other and provide better performance when used together.

In RST, given a text document, a graphical structure/tree structure is constructed, where nodes represent texts and edges represent the discourse relations. The text connected by discourse relations falls into two categories namely Nucleus and Satellites. Nucleus expresses the salient part of the text, whereas the satellite contains the additional information about the nucleus. The Nucleus, Satellite and the discourse relation form the smallest discourse unit known as Elementary Discourse Unit (EDU). We denote the pattern or the structure formed by a nucleus, satellite and the discourse relation, as NRS sequence in this paper. Figure 1 shows the Nucleus, Satellite and the discourse relation connecting them (NRS sequence) for an English sentence given in Example 1.

Example 1: If you come to my school tomorrow, you can meet my teacher

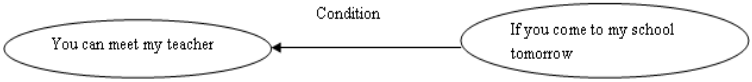

FIGURE 1- NRS sequence for Example 1

On the other hand, sangatis are defined as the content that induces the desire to know what is being said next in text (Madhava Charya, 1989)**.** Sangatis are typically used in the explanation of *sūtra* -based texts. *sūtr*as express content in crisp, short statements; *adhikaraṇa* (sub-topic) is the organization of a set of related *sūtra*s. A set of *adhikaraṇa*s form a *pāda* (section), and a set of *pāda*s form an *adhyāya* (chapter). *sūtra*s being cryptic in nature need to be explained. The explanation is normally organized at the level of *adhikaraṇa*. An *adhikaraṇa* is said to have five components, namely, subject of discussion, doubt/ambiguity in understanding the subject, sangati for this discussion, opponent's view and the proponent's (proposed) view. Of these, sangati is explained at various levels. At the *sūtra* level in terms of how this *sūtra* is related to the previous *sūtra*; at the *adhikaraṇa* level as to how this *sūtra* is relevant to the *adhikaraṇa*, at the *pāda* level as to how it is relevant to that *pāda* and so on. Similarly, sangati is discussed between *adhikaraṇa*s, and between *pāda*s as well. Examples of sangatis include *upodghāta*, *apavāda*, *ākṣepa* and *prāsan*gika. Figure 2 illustrates, upodgata sangati describing text on cancer. *upodghāta* links the introductory part of any text, to its respective explanatory part of the text.

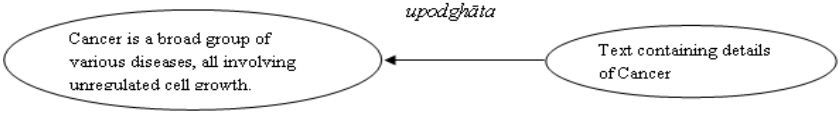

FIGURE 2-Usage of *upodghāta* sangati

This paper is organized as follows. Section 2 discusses the work related to discourse parsing. Section 3 illustrates the power of sangatis in representing text and about the proposed sangati based discourse parser; explains the comparison of RST and sangati and describes the RST-Sangati based discourse parser. Section 4 provides the evaluation of the discourse parsers. Conclusions and reference section follow section 4.

2 Related Work

Many techniques have been proposed for discourse parsing. Marcu et have come up with an unsupervised approach that identifies five discourse relations namely, Contrast, Explanation, Evidence, Condition and Elaboration al (Marcu et al, 2002). They have used discourse markers and frequently co-occurring word pairs to identify the discourse relations. Hassan et al have designed a rule based discourse parser for Arabic language and they have used cues to identify the discourse relations (Hassan et al, 2008). Using cue phrases, many discourse parsing techniques have been proposed in various languages namely, Mandarin (Songren, 1985), Spanish (Lorraine, 1986), Thai (Sithipoun et al, 2010), and Bengali (Dipankar Das et al, 2010). Recently, Hugo Hernault et al have come up with a discourse parser named HILDA that labels discourse relations using Support Vector Machines (SVM) (Hernault et al, 2010). It has also been claimed that HILDA is computationally more efficient than the earlier techniques on discourse parsing proposed by Reitter (Reitter, 2003) Baldridge et al (Baldridge et al, 2005). HILDA uses lexical and syntactic features trained by the SVM classifier. SVM has been used both as a binary classifier as well as a multi class classifier for discourse parsing.

As stated previously, the discourse parser discussed in this paper is the extension of our previous work on language independent discourse structure framework using Universal Networking Language (UNL) (Subalalitha et al, 2011). UNL is a computer language that enables computers to process information and knowledge across the language barriers. Given a sentence, UNL represents it as graph, with nodes as concepts and UNL relations as edges which is known as Enconversion (Enconversion Specifications, 2009). There are 46 UNL relations identified by UNDL. Obj (Object), agt (Agent), plc (Place) are few UNL relations. For instance, for the sentence shown in Example 2, the agent concept, "Ram *(iof>person)"* and the verb concept, "kill (icl>action)*"* are connected by the UNL relation, "agt" and the object concept, "Ravana (iof>person)" is connected to the verb concept by the UNL relation, "*obj"*.

Example 2: Ram killed Ravana

In the existing RST based discourse parser, the NRS sequences are identified by exploring the similarities that exist between UNL relations and discourse relations relations (Subalalitha et al, 2011). Also by using UNL, the discourse structure formed, becomes language independent which is first of its kind. This paper proposes a similar language independent discourse parser that identifies sangatis using UNL. The details of sangatis and how they are used to build a discourse structure is discussed in the next section.

3. RST-Sangati: A comparison:

While comparing RST based discourse relations with sangati, both representations are unique in their own way. Table 1 shows the list of sangatis considered along with their English meaning and explanation. Also, discourse relations that are similar and equivalent are listed. Like discourse relations certain sangatis are multi nuclear which is mentioned in the table.

S.No	Sangati	Equivalent and Similar Discourse Relations		
		Equivalent Discourse Relation	**Similar Discourse Relation**	**Explanation**
1	Up*odghāta* (Introduction)	Preparation	-	*upodghāta* sangati gives an introduction of the text. Preparation discourse relation prepares the reader to expect and interpret the text to be presented.
2	*apavāda* (Exception) (multi nuclear)	-	Contrast	*apavāda* sangati indicates an exceptional Scenario. Contrast discourse relation may indicate an exceptional scenario but not always
3	*ākṣepa* (Objection) (multi nuclear)	-	Contrast	*ākṣepa* sangati indicates an objectional statement. Contrast discourse relation may indicate an objectionable statement but not always
4	*prāsan*gika (Related) (multi nuclear)	-	-	*prāsan*gika is a unique sangati and does not have an equivalent or similar discourse relation
5	utt*āna* (Arises) (multi nuclear)	-	Antithesis	utt*ān*a sangati indicates a new issue related to the text. Antithesis discourse relation may indicate a new issue but not always
6	*sthirīkaraṇa* (Strengthen)	-	Justification	*sthirīkaraṇa* sangati indicates a strengthening text supporting the topic in focus new issue related to the text. Justification discourse relation may indicate a supporting reasons but the supporting reason may not be a strengthening reasoning always
7	*ātidesîka* (Transference) (multi nuclear)	-	-	*ātidesîka* is a unique sangati and does not have a similar or equivalent discourse relation.
8	pratyavasthana (Re-instate)	-	-	pratyavasthana is a unique sangati and does not have a similar or equivalent discourse relation.

9	*dṛṣṭanta* (Example)	-	Elaboration	*dṛṣṭanta* sangati indicates an example Scenario Elaboration discourse relation may indicate an example scenario but not always.
10	*pratyudharaṇa* (Counter Example)	-	Contrast	*pratyudharaṇa* sangati indicates an counter example scenario Contrast discourse relation may indicate a counter example scenario but not always.
11	Anantara (Follows) (multi nuclear)	Sequence	-	anantara sangati links the texts in sequence similar to the Sequence discourse relation
12	*viśeṣa* (Special Case)	-	Elaboration	*viśeṣa* sangati explains a speciality. Elaboration discourse relation may describe a specialty but not always.

TABLE 1- **A comparison between sangatis and Discourse Relations**

It can be observed from the Table 1 that, most of sangatis are unique and specific to a scenario. A discourse relation can be mapped to one more scenarios. For instance, the scenario linked by an Elaboration discourse relation may be an explanation, additional information or it may be a strengthening statement to the scenario. Whereas, distinct sangatis are available to handle the different scenarios handled by the Elaboration discourse relation. For instance, if the additional information of a scenario is just an added information but related to the scenario, it can be linked by *prāsaṅgi*ka sangathi. If the explanatory text denotes an example or a counter example, it can be linked by *dṛṣṭanta* and *pratyudharaṇa* sangatis. If the explanatory text describes the speciality of the scenario, it can be linked by viśeṣa sangati. On the other side, there are unique RST based discourse relations as well. This paper considers the list of RST based discourse relations considered by Mann et al (1988). It is observed that, there are many discourse relations that handle various scenarios which is not possible with sangatis. Concession, Evaluation, Background are some of the unique discourse relations. Also apart from the list of sangatis discussed here, there are more number of sangatis that need to be explored. For clausal level discourse analysis, discourse relations are more effective than sangatis. Whereas, sangatis go well beyond clause level. So for an efficient and complete discourse parsing, discourse relations and sangatis need to be used in union. The next section discusses how these sangatis are identified.

3. 2 Identifying Sangatis:

The sangati based discourse parser makes use of only UNL to identify the sangatis whereas, the RST-Sangati based discourse parser makes use of both UNL and RST based discourse relations to identify sangatis. This section illustrates about the sangati based discourse parser.

3.2.1 Sangati Based Discourse Parser:

The UNL components such as semantic constraints and UNL relations are used in combination to identify the sangatis. This is similar to the identification of RST based discourse relations done in the existing work (Subalalitha et al, 2011). For example, in the sentences given

in the Example 3, the *prāsaṅgika* sangathi is identified from the UNL relations and UNL concept similarity that exists between the UNL graphs constructed for each sentence.

Example 3: I went to Chennai last Friday. Travelling to Chennai during Fridays is terrific.

The rules for identifying sangatis using UNL components are given in Table 2. The rules provide the seed feature set for seven sangatis. The additional features for each sangati are learnt using Naive Bayes Probabilistic classifier whose details is given in the next section.

S.No.	Sangati	Rules
1	*upajīvya*	Presence of iof, nam and met UNL relations in Satellite along with UNL concept/ semantic constraint similarity between Nucleus and Satellite UNL graphs
2	*ātidesîka*	UNL graph similarity between the nuclei in terms of concepts, except the main subject.
3	anantara	Presence of Seq UNL relation in one of the nuclei UNL graphs. Also used as default relation for tourism documents.
4	*viśeṣa*	Presence of Cue such as "speciality", "uniqueness" in satellite UNL graph+ UNL relation "pos" in nucleus UNL graph.
5	*prāsaṅgika*	Presence of identical UNL relations and with UNL concept/ semantic constraint similarity between two nuclei UNL graphs.
6	*upodghāta*	Presence of UNL concept that is connected to the verb frequently in the document UNL graph which becomes the nucleus and the rest of the text becomes the satellite.
7	*dṛṣṭanta*	Presence of iof UNL relation in the UNL graphs in the satellite UNL graphs along with UNL concept/ semantic constraint similarity between the nucleus and satellite UNL graphs.

TABLE 2- **Rules for identifying sangatis using UNL**

3.2.2 Learning with Naïve Bayes Probabilistic classification

For a set of text documents, sangati representation of the texts is constructed using the seed features listed in Table 2, which forms the training set. Learning of new features using Naive bayes probabilistic classifier is discussed below. For a sangati S_i, the nuclei and satellites connected by it are used as the context. Let S_{edu} denote the set of sangatis listed in table 2.

a) For each sangati $S_i \epsilon S_{edu}$, all the nuclei sub graphs connected by S_i are extracted from the training set.

b) The nuclei sub graphs extracted at step 1 contains a set of UNL relations in them. These UNL relations are considered as the context window from which the additional features could be learnt. For instance, if m number of nuclei sub graphs is extracted for sangati S_i, m number of series of UNL relations is extracted. These series indicate possible additional features that could signal sangati S_i apart from the seed feature(s) listed in table 2.

c) Let us denote the additional features as f_i, where i ranges from 0 to m. Bayesian probabilities $P(f_i/S_i)$ are computed for all features.

d) Total Probability, $Prob_{tot} = \sum_{i=0}^{i=m} P(fi/Si)$ and the average probability , $Prob_{avg} = Prob_{tot}/m$ are calculated.

e) The features whose probabilities are more than the $Prob_{avg}$ are chosen as additional features that could signal the sangati S_i.

f) In the testing phase, the classifier tries to match the features present in the nuclei sub graphs with the seed features and the additional features learnt and identify the sangati.

3.3 RST-Sangati Based Discourse Parser

The RST-Sangati based discourse parser identifies both sangatis and RST based discourse relations using UNL and RST. The rules for identifying sangatis using UNL and RST are given in Table 3.

S No	Sangati	Rules
1	*upajīvya*	Presence of Elaboration and Means or Re-instate discourse relations in the satellite UNL graph along with UNL concept/ semantic constraint similarity between the nucleus and satellite UNL graphs.
2	utt*āna*	Presence of Contrast and Antithesis discourse relations in NRS sequence along with UNL concept/ semantic constraint similarity.
3	*ātidesîka*	UNL graph similarity in terms of concepts, except the main subject concept between the nuclei UNL graphs.
4	anantara	Presence of Sequence discourse relation in one of the nucleus UNL graph. Also used as default relation for tourism documents.
5	*viśeṣa*	Presence of Cue such as "speciality", "uniqueness" in satellite UNL graph + UNL relation "pos" in nucleus UNL graph.
6	*prāsaṅgika*	Presence of identical UNL/Discourse relations between the nuclei UNL graphs.
7	*upodghāta*	Presence of UNL concept that is connected to the verb frequently in the document UNL graph which becomes the nucleus UNL graph and rest of the text becomes the satellite.
8	d*ṛṣṭanta*	Presence of Elaboration discourse relation in the satellite UNL graph along with UNL concept/ semantic constraint similarity between the nucleus and satellite UNL graphs.

TABLE 3- **Rules for identifying sangatis using RST and UNL**

Since the input documents are tourism domain specific, the rules also lean towards the same domain. Like sangati based discourse parser, these rules are used as seed feature set and more features for each sangati are learnt using Naive Bayes probabilistic learning. Finally, a language independent discourse structure using discourse relations and sangatis is given as output, which can be used as background for many NLP applications.

4 Evaluation

We have tested our sangati based discourse parser and RST-Sangati based discourse parser using 500 Tamil tourism domain documents as training data and freely available 21 RST-Discourse Tree (RST-DT) files as test data. By using both Tamil and English text documents, we have shown the language independent quality of our discourse parsers. We have also made a comparison with RST based discourse parser. Table 4 lists the Precision, Recall and F-score of the RST, sangati and RST-Sangati discourse parsers.

Factors	RST based discourse parser	Sangati based discourse parser	RST-Sangati based parser.
Precision	91.19%	74.62%	96%
Recall	68%	57.3%	87.99%
F-score	79%	64.84%	93.36%

TABLE 4 -Precision, Recall and F-score of RST, sangati and RST-Sangati discourse parser

It can be seen that, the RST –Sangati based parser shows higher precision, recall and f-score values compared to RST based discourse parser and sangati based discourse parser. This is due to the fact that in RST based discourse parser, the complex coherent relations beyond clause level is not captured b the RST based discourse relations and similarly in sangati based discourse parser, the simple coherent relations at clausal and sentence level is not captured by sangatis . Also, it can be observed that the RST based discourse parser shows better performance than sangati based discourse parser. This is due to the expository texts used as the corpus where more RST based discourse relations are observed than sangatis. It can be seen that, the proposed RST-Sangati based discourse parser identifies only eight sangatis that are specific to tourism domain. Also, the proposed discourse parser provides a very basic text representation model which needs to be enhanced by analyzing various genres of texts containing arguments and stories.

Conclusions

A discourse parser that identifies NRS sequences based on RST and sangati has been presented in this paper. sangatis as per Sanskrit literature expresses continuity between *sūtra* based texts. Sangatis are similar to discourse relations that connect coherent parts of the text. sangatis can connect more complex discourse units which is difficult with discourse relations. We have proposed two discourse parsers namely sangati based discourse parser which identifies only sangatis and a RST-Sangati based discourse parser which identifies both RST based discourse relations and sangatis. The sangati based discourse parsers uses only UNL to identify sangatis whereas, the RST-Sangati based discourse parser uses both UNL and discourse relations to identify sangatis. It is observed that RST and sangati complement each other well so better performance is obtained when they are used together than when used alone. The discourse structure built by the proposed discourse parser can be a back bone for many NLP applications such as QA, IR and IE systems.

References

J. Baldridge and A. Lascarides (2005):. Probabilistic head-driven parsing for discourse structure. In Proceedings of the Ninth Conference on Computational Natural Language Learning, 2005 volume 96, page 103.

Cui, Songren (1985).. Comparing Structures of Essays in Chinese and English Mastres Thesis, U.C.L.A.

Daniel Marcu and Abdessamad Echiha (2002). . An Unsupervised Approach to Recognizing Discourse Relations In Computational Linguistics (ACL), Philadelphia, pp. 368-375.

Dipankar Das, Sivaji Bandyopadhyay (2010). Labeling Emotion in Bengali Blog Corpus-A fine Grained Tagging at Sentence Level In Proceedings of the 8th Workshop on Asian Language Resources, pages 47–55.

Enconverter Specifications, UNL center, UNDL Foundation, 2009.

Hassan I. Mathkour, Ameur A. Touir and Waleed A. Al-Sanea (2008). Parsing Arabic Texts Using Rhetorical Structure Theory. In Journal of Computer Science 4 (9): 713-720.

Hugo Hernault, Helmut Prendinger. David A. duVerle, Mitsuru Ishizuka (2010).HILDA: A Discourse Parser Using Support Vector Machine Classification (2010). In Dialogue and Discourse 1(3) 1-33 doi: 10.5087/dad.2010.003.

Kump Lorraine. Structuring Narrative text in a Second Language (1986). descriptive of Rhetoric and Grammar Phd Thesis,University of California Los Angeles.
Madhava Charya (1989). Jaiminiya Nyaya Mala Vistara. Chankhamba Sanskrit Pratishthan.

Mann, W.C., & Thompson, S.A (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. In Text (3). 243-281.

D. Reitter (2003).Rhetorical Analysis with Rich-Feature Support Vector Models. Unpublished Master's thesis, University of Potsdam, Potsdam, Germany , 2003.

Somnuk Sithipoun , Om sornil (2010). Thai Rhetorical Structure Analysis. International Journal of Computer Science and Information Security,Vol. 7, No. 1.

Subalalitha C.N and Ranjani Parthasarathi (2011). A Language Independent Rhetorical Structure Framework Using Universal Networking Language. Proceedings of fifth Indian International Conference on Artificial Intelligence, 2011 page no-1427-1440.

The Universal Networking Language (UNL) Specifications (2009), UNL center, UNDL Foundation.