CHAPTER 1

INTRODUCTION

Natural Language Processing (NLP) involves processing and analyzing the information present in a Natural Language (NL) text. The ultimate goal is to make the NL text understandable, which in turn, requires its meaning to be captured. An NLP task can be done at various levels, namely, morphology, syntax, semantics, pragmatics, and discourse. Morphology deals with the formation of words from the basic unit, “Morphemes”; syntax is related with forming a grammatically correct clause or sentence, using the grammatical role of the words; semantics is concerned about the meaning of the words and the sentences formed by the words; pragmatics focusses on how the interpretation of a text is affected due to the usage of words in different contexts, and discourse is concerned about interpreting one text along with the related texts. All these NLP tasks handle the text at different levels, namely, word, clause, sentence, paragraph and document. This thesis proposes a text representation framework named, “SemTRex” which aims at constructing semantic structures at various levels, namely, word, clause, sentence, paragraph and document. SemTRex aims at processing the text at the discourse level.

The Rhetorical Structure Theory (RST) is a widely used theory to analyse the text at the discourse level and to form a discourse structure representation called, rhetorical structure (Mann & Thompson 1988). The SemTRex framework proposes a semantic based language independent approach for the construction of a discourse structure, and enhances the
discourse structure by constructing a higher level concise representation on top of it. The idea behind forming such a representation, is to capture the semantic essence of the discourse structure which can be seen as a semantic index to it. The semantic index can aid NLP applications to benefit from the discourse structure information. The construction of the rhetorical structure requires word-level semantics, which is usually supplied by lexical resources such as, WordNet, Ontology, Frame Net etc. (Marcu et al 2002). The SemTRex framework puts forth an alternative semantic representation, namely, the Universal Networking Language (UNL) a language independent framework to provide the semantics for constructing the discourse structure.

1.1 RESEARCH MOTIVATION

The idea of a concise representation of the discourse structure was inspired by the concept called, “śūṭra”, which has been used in ancient Tamil and Sanskrit language literatures, to express large content in a crisp manner (Pavananthi Munivar 1994, Macdonell 1900). In Nannool, a Tamil grammar masterpiece, śūṭrās have been used to define the grammar rules. In Sanskrit literature, śūṭrās have been used in many śāstrās – grammar, nyāya, Veḍhanṭha etc. śūṭrās are like short semantic representations of text documents.

As per both the literatures, śūṭrās can be related to formulae that represent a large piece of text in a brief and crisp manner. This idea of the śūtra is equivalent to semantic indexing in current-day document processing. The semantic index of a document needs to capture the semantic essence of the document, so that NLP applications can access the content of the document through the semantic index. This idea of semantic indexing is similar to how the śūtra was used, to capture the semantic essence of a text in Tamil and Sanskrit literatures. Hence, the proposed research makes use of this idea in building a concise representation of the discourse structure, which is
inherent with the semantic relations linking the text units present in a document.

To interpret the sūtras, a concept called, “saṅgatis” has been used in Sanskrit literature. saṅgatis have been used in sūtra based texts to link and provide continuity between the sūtras. This idea of saṅgatis can be compared with the RST based semantic relations that capture the coherence or continuity between the text units. Inspired by the qualities of sūtras and saṅgatis and their similarities with the modern text processing techniques, the proposed research aims at constructing a concise semantic representation like the sūtra, which can serve as a semantic index to the discourse structures built using RST and saṅgatis. By integrating the ancient and the modern techniques, the SemTRex framework attempts to acquire the benefits from all of these techniques, thereby building efficient semantic structures.

The definition and explanation of sūtras and saṅgatis are given in the next section.

1.1.1 sūtra

The definition of the sūtra as per Nannool is given below in Tamil followed by the English transliteration and its meaning in English.

Definition of sūtra

\[
\text{சிறுமை} \\
\text{செம்மை} \\
\text{வாழ்ச்சூர்}
\]
English Transliteration:

Cil vakai eluttīl palvakaip porulai
cevva nāṭiyir čeṛttinītu viḷakkit
tīṭpa nuṭpaṇ cirantaṇa čūttiram

Gloss:

The speciality of the śūtras is to coherently convey the semantics precisely, accurately, and with certainty, using a few words.

As per Sanskrit Literature the definition of śūtra is transliterated in English as follows.

alpākṣaraṃ asandigdhaṃ śūravad viśvatomukham astobhaṃ
anavadyaṃ ca śūtram śūtravido viduh.

The meaning of the above definition is, “Of minimal syllabary, unambiguous, pithy, comprehensive, continuous, and without flaw: who knows the śūtra knows it to be thus.”

As per Sanskrit literature, śūtra is an aphorism (or line, rule, formula) or a collection of aphorisms in the form of a manual or, more broadly, a text in Hinduism or Buddhism. Literally, it means a thread or line that holds things together and is derived from the verbal root sīv-, meaning, “to sew”.

It can be observed that the śūrās have been used by both the literatures to express the content of a coherent text in a concise manner. The first line in the definition of the śūrās as per Nānool, “Cil vakai eluttīl palvakaip porulai” (representing multiple semantics in a few words) denotes the characteristic of a text index. An efficient text index should be the
representative of a sentence, paragraph, or document from which it is extracted. Furthermore, śūtrās express continuity between the texts they represent. Hence, inspired by these qualities of śūtrās, this thesis proposes a semantic text representation technique for discourse structures.

1.1.2 saṅgati

sangatis are defined as the content that induces the desire to know what is being said next in a text (Madhava Charya 1989, Amba Kulakarni & Monali Daas 2012 and Monali Daas & Amba Kulakarni 2013). saṅgatis are typically used in the explanation of śūtra-based texts. As per Sanskrit literature, the components of a text are divided into, śūtrās, adhikaraṇa (sub-topic), pāda (section) and adhyāya (chapter). A set of śūtrās form an adhikaraṇa; a set of adhikaraṇas form a pāda; a set of pāda form an adhyāya. saṅgatis provide the link between these components along with the explanation.

Examples of saṅgatis include upodghāta (Introduction), apavāda (Exception), ākṣepa (Objection) and prāsangika (Related). The proposed research focusses on twelve such saṅgatis for the discourse structure construction.

The next section discusses the RST, which is similar to saṅgatis, in terms of capturing the continuity and used in current-day document processing.

1.2 RST: STATE-OF-ART

RST is one of the theories of discourse, proposed by Bill Mann, Sandy Thompson and Christian Matthiessen, of the University of Southern California, as part of the studies of computer-based text generation (Mann &
Thompson 1988). The text units, namely, clauses, sentences and paragraphs are interwoven with one another with some hidden semantic relations between them. When these relations are identified, the coherence or the continuity existing between the texts can be brought out. RST captures the coherence between the text using discourse relations. For a given text document, the RST builds a graph or tree like discourse structure, where nodes represent the texts, and edges represent the discourse relations. The discourse relations fall into two categories, namely, nucleus and satellite. Nucleus expresses the salient part of the text, whereas the satellite expresses the additional information about the nucleus. A text unit represented by a nucleus or satellite can be a clause, sentence, paragraph or a document. Evaluation, Circumstance, Means and Condition are some of the discourse relations used in RST.

NLP applications can make use of a discourse structure as such, or build an index which can aid query based NLP applications, such as Information Retrieval (IR) and Question Answering (QA) systems. Many researchers have focussed on these perspectives. In order to bring out the importance of the proposed research, the state-of-art RST is discussed in the following sections.

1.2.1 RST-Based Discourse Parsing

Many research improvements have been made in the past two decades, in terms of various factors involved in discourse parsing, such as, features used, level of text processing (sentence/paragraph/document), machine learning algorithms and accuracy of the whole process.

One of the early attempts in discourse parsing was constructing unambiguous RS trees, given a text (Marcu 1996). Most of the discourse parsing-related works have identified discourse relations between clauses and
sentences using cue phrases, lexical and syntactic features. Learning algorithms such as, Naive Bayes classification have also been used to identify discourse relations (Marcu & Echihabi 2002). A sentence level discourse parsing algorithm called SPADE (Sentence-level Parsing for Discourse) uses lexical and syntactic features, to identify discourse relations even in the absence of cue phrases (Soricut & Marcu 2003). A discourse parser named, Linguistic Discourse Analysis System (LIDAS) uses sentential syntax and lexical semantics to build an RS Tree for a given NL clause or sentence (Polanyi et al 2004). A first-order logic learning approach uses linguistic cues and lexical semantics to identify discourse relation between clauses (Subba & Di Eugenio 2009).

A few discourse parsers have been developed to identify discourse relations between paragraphs. A discourse parser named, High-Level Discourse Analyzer (HILDA) uses shallow lexical and syntactic features, to identify discourse relations between sentences and paragraphs (Hernault et al 2010). HILDA uses Support Vector Machines (SVM) to identify the discourse units and to label the discourse relations. Surface features (Part Of Speech (POS) tag, trigram), syntactic features (tense, aspect and polarity), lexical features (cue phrases), reference features (pronouns) and discourse features (punctuation marks) have been used in union to identify the preceding and the following paragraphs (Theijssen et al 2008). Furthermore, five machine learning algorithms namely, Naive Bayes, k-Nearest Neighbours (kNN), SVM, Decision Trees and Maximum Entropy have been used to extract these features from the corpus.

Cue phrases, contextual features, discourse production rules, and semantic similarities have been used to construct a document level discourse parser (Feng & Hirst 2012). This work has used an improved set of features by considering the knowledge of the discourse relations of the preceding and
the following text units for discourse relation labelling, in addition to the normally used feature set for discourse parsing.

Apart from English, discourse parsers have also been constructed for various other languages, such as, Arabic and Thai, using cue phrases and syntactic features (Hassan et al 2008 and Sinthupoun & Sornil 2010).

1.2.2 **RST Based Indexing Techniques for NLP Applications**

Many RST based NLP applications incorporated with computational mechanisms have been developed (Taboada & Mann 2006). Its ability to capture the coherence of an NL text aids the NLP applications to acquire a better understanding of the text. The conceptual discourse structure and the discourse relations that map the coherence inherent in the text enable NLP applications to access the hidden semantics in the NL text. A single rhetorical structure can cater to various NLP applications such as, IR, summary and QA systems. The cue words present in the user queries of these NLP applications find a direct mapping with the discourse relations, which is one of the most important reasons behind the success of the RST based NLP applications. Some of the existing works on RST based NLP applications are discussed in this section.

IR systems have used RST for semantically indexing the documents by considering the discourse relation between the text fragments (Haouam & Marir 2006 and Sahib & Ali shah 2006). Both these discourse parsers have used cue phrases to identify the discourse relations between the clauses. In both these techniques, the RST components are indexed as the triplet, \(<RelationName, Nucleus, Satellite>\) and stored in the data base, and a weight factor is given to each index based on its frequency of usage, and the number of important words present in it.
RST has also been used to shorten the texts by extracting the important components from the discourse structures, which have been used in summary, and abstract generation related works. A technique called, “Variable Length Documents (VLD)”, generates a summary by choosing the important part of the text, using the Rhetorical structures based on the weights assigned to each RST component (O'Donnel 1997).

A RST based query focussed summary generation system has been developed by Bosma (2004). The discourse structure is represented as a graph. A QA system is needed by this summary generation system to locate a node from the discourse graph that is relevant to the query. The summary generation system will append the nodes that are linked to this node in the graph, based on some weight factors to form a summary. Louis et al (2010) have attempted to find the important sentences to form a summary, which is similar to the approach proposed by Bosma (2004).

Rhetorical structures have also been used for abstract generation and separation of the topic and non-topic parts from the text. (Ono et al 1994, Shibata & Kurohashi 2005).

An RST based indexing technique has also been used in the development of QA systems (Verbene et al 2007). This technique is mainly designed to answer, “why” type of questions using probabilities that involve the nucleus and satellites with respect to a pre-defined question set. The assumption is that, the nucleus contains the question and the respective satellite contains the answer.

It can be observed that the existing RST based discourse parsing techniques attempt to build discourse structures for an NL text, using various features, namely, lexical, syntactic, semantic, reference, and discourse features. The discourse relations have been identified between the text units,
namely, clause, sentence, paragraph and document levels. It can be observed that most of the existing works have used cue phrases, and lexical and syntactic features for discourse parsing. The content of an NL text can contain words with various senses and meanings. Lexical and syntactic features may not be sufficient to interpret the text units using discourse relations that are not explicitly connected by cue phrases. Moreover, all the existing discourse parsers aim at building a discourse structure for a specific language, and language independence still remains a challenge.

From the discussions done on the NLP applications built using the RST, it can be observed that the advancements that have happened in the discourse parsing arena, in terms of using rich features for identifying the nucleus, satellite and the discourse relations, is yet to reflect on the applications. Most of the RST based NLP applications such as IR, QA and summary generation system use a cue phrase based discourse parser, and they are capable of identifying discourse relations only at the clause and sentence levels. The complete advantage of a discourse based NLP application can be acquired, only when the discourse parser can handle a text beyond the sentence level. Capturing the relations between paragraphs and documents aids in developing NL applications that can achieve accuracy even while handling complex queries.

1.3 SemTRex FRAMEWORK

1.3.1 Research Focus

The SemTRex framework aims at constructing semantic structures and an index at various levels, namely, word, clause, sentence, paragraph and document. The UNL has been used to explore the word level semantics, while the RST and saṅgatis have been used to explore the semantics beyond the word level. Finally, a crisp representation of the semantic structure based on the ideas of sūtra is built. Construction of semantic structures at higher levels
makes use of the semantics at the lower levels. For instance, the UNL based word level semantics is used while constructing the RST based discourse structure at the clause, sentence, paragraph and document levels; the saṅgati based discourse structure makes use of the semantics of both UNL and RST, and the final sūtra form makes use of the UNL, RST and saṅgatis.

In order to explore the word level semantics using the UNL, the SemTRex framework constructs a UNL indexer, which identifies UNL based indices from a UNL representation. The UNL captures the semantic relations between the words. The words are converted to language independent concepts and are linked to the other concepts through semantic UNL relations. This conversion is known as, “Enconversion” (Uchida et al 1999). Given an NL clause or sentence, UNL represents it as a graph, with concepts as nodes and UNL relations as edges. There are totally 53 UNL relations identified by UNDL. \textit{Obj} (Object), \textit{agt} (Agent), \textit{plc} (Place) are some of the UNL relations. The UNL indices may be considered as a first-level version of sūtra representation which uses only word level semantics.

The similarity of UNL and RST in terms of the semantic relations and di graph nature is explored, to scale up the UNL based word level semantics to the higher levels. This is done by constructing RST based discourse structures using UNL based features. Since UNL is a language independent framework, the RST based discourse structure which is built using UNL also becomes language independent. The UNL-RST discourse parser makes use of UNL based features in all the stages of discourse parsing namely, segmentation, nucleus-satellite identification and discourse relation labelling. An initial feature set is framed and Naive Bayes learning is used to learn additional features.

Having built a language independent UNL-RST discourse structure, the SemTRex framework proceeds further to construct UNL-saṅgati and
UNL-RST-saṅgati discourse structures. The former makes use of only UNL based features while the latter makes use of both UNL and RST based features. It is observed that UNL relations, RST relations and saṅgatis have certain qualities in common, which are explored to construct a discourse structure using these techniques in union. The UNL-saṅgati discourse parser is constructed as an initial attempt to test the suitability of saṅgatis for discourse parsing. The UNL-saṅgati discourse parser handles texts at the clause or inter-sentence, and sentence levels, while the UNL-RST-saṅgati discourse parser handles texts at the clause, sentence, paragraph and document levels. Naive Bayes learning is incorporated in the UNL-saṅgati discourse parser, while the UNL-RST-saṅgati discourse parser is a rule based discourse parser.

The SemTRex framework finally builds a concise representation for the UNL-RST-saṅgati discourse structure. The sūtra representation captures the semantic essence of UNL, RST and saṅgatis. The sūtra representation can be viewed from two perspectives, namely, text representation and semantic indices. The SemTRex framework has used the sūtra representation as semantic indices in two NLP applications, namely, the IR system and summary generation systems.

1.3.2 Research Contributions

Since the sūtra representation serves both as a text representation and semantic indices, the contributions of this research can also be viewed from these two perspectives. The following are the contributions from the text representation point of view.

- Construction of a language independent UNL-RST discourse structure.
Using UNL- RST based discourse structure frame work for Tamil thereby providing the first RST solution for Tamil.

Usage of saṅgatis for discourse parsing.

Construction of language independent UNL-RST-saṅgati discourse parsers

sūtra based concise semantic text representation for a discourse structure

From the semantic indexing point of view, the contributions of the SemTRex framework are as follows.

Construction of UNL indexer which forms the basis for identifying the concise semantic text representation at the word level.

Construction of sūtra based indexer for the RST-saṅgati based discourse structure

Usage of sūtra based indices in constructing two NLP applications namely, IR and summary generation systems.

1.4 OUTLINE OF THE THESIS

The thesis is organized as follows. Chapter 1 gives an introduction to the thesis. Chapter 2 discusses the various semantic text representation, and indexing techniques. Chapter 3 provides details about the overview of the SemTRex framework and about the UNL indexer. Chapter 4 describes the language independent UNL-RST discourse parser. Chapter 5 compares RST and saṅgati. Chapter 6 is about the UNL-RST-saṅgati discourse parser. Chapter 7 discusses the sūtra based text representation and its usage in NLP applications namely, the IR system and summary generation system. Chapter 8 gives the conclusion of the thesis.