Annotations of Connectives and Arguments in Malayalam Language

Kumari Sheeja S, Sobha Lalitha Devi*

* AU-KBC Research Centre, Anna University, Chrompet, Chennai,
KCG College of Technology, Karapakkam, Chennai,
sobha@au-kbc.org

Abstract

Discourse relations in natural languages link clauses in text and compose overall text structure. Discourse connectives are an important part of modeling the Malayalam discourse structure. We followed the annotation procedure of Penn Discourse Tree Bank and worked on tagging of discourse connectives and arguments of Malayalam text and also report the senses of relation. We present our work on annotations of Malayalam discourse connectives and arguments which helps to know more about the discourse connectives and their appearance in case of semantic rules in Malayalam discourse. Discourse connectives may or may not be explicitly present in the relation. In our work, we focus on the annotation of both explicit and implicit connectives and arguments in Malayalam text and showed encouraging results.

Keywords: Malayalam discourse; discourse connectives; annotation; explicit; implicit connective

1. Introduction

Natural Language Processing is a field of Computational Linguistic which is mainly used for various text processing in natural languages such as part-of-speech tagging, word segmentation, word sense disambiguation, named entity recognition, parsing etc. Discourse analysis is one of the most important task in Natural Language Processing that has many applications in natural languages such as text summarization, information extraction, opinion mining etc. Discourse connectives are extensively used in NLP applications to describe how two sentences or clauses are semantically connected in large text. Discourse connectives may or may not explicitly exist in the text. Explicit connectives signal the presence of discourse connectives between sentences or clauses.

Ramu is an intelligent boy but he cannot score good marks in the examination.

* Corresponding author. Tel.: +0-91984-013-3444;
E-mail address: sobha@au-kbc.org
In this example the connective “but” makes a relation between two clauses or sentences and making the text coherent. In implicit discourse relation, there is no connective in between the arguments of text which results a very difficult task. We describe an approach to annotate a large scale corpus in terms of more basic characterization of discourse structure in Malayalam text in terms of discourse connectives and their arguments.

Discourse connectives have multiple senses like verbs depends on the semantic structure of the text in Malayalam discourse. In this paper our effort to annotate Malayalam discourse connectives and arguments based on the guidelines of PDTB. The content of the paper started with introduction of the work. Section 2 describes the literature survey of our paper. Section 3 gives an overview of discourse relations in Malayalam, section 4 describes Malayalam text corpus and annotation process and section 5 describes the sense annotations of connectives. Finally the paper ends with the conclusion of the work.

2. Related Work

Annotation process of discourse connectives and their arguments have been explored in various languages such as Hindi[1], Tamil[3], Arabic[10], etc. They worked on tagging of discourse connectives and their arguments. They developed discourse relation bank of Hindi, Tamil and Arabic that will be useful as a resource for further researches in discourse. Versely[5] worked on tagging German discourse connectives and arguments using English training data and a German–English parallel corpus. Versely’s approaches were to transfer a tagger for English discourse connectives. They have done this work using a freely accessible list of connectives by annotation projection. Annotating the discourse connectives in Turkish language is done by Zeyrek, D., and Webber, B. L. [7]. Faiz, S. I., and Mercer, R. E. [9] worked on explicit connectives and their arguments in discourse with examples. Wang et al.,[6] used sub-trees as features and achieved a significant improvement in identifying arguments, explicit and implicit discourse relation. Yuping Zhou and Nianwen Xue[11] worked on tagging the discourse relations and their senses in Chinese language. In this paper we have explored the annotation results of various discourse connectives and their arguments in Malayalam language.

3. Discourse Connectives in Malayalam

Malayalam, a South Indian or Dravidian language is a free word order language but maintains the verb in final position. Discourse connectives in Malayalam text are important for interpreting or producing text. Malayalam has a special place in the classification of languages. There are different spoken forms in Malayalam even though the literary dialect throughout Kerala is almost same. Malayalam has its own distinct script and most of the syntactic information is embedded as morphological structure. This causes the analysis of word forms of Malayalam to cross the limits of morphology and it reaches to syntactic and semantic level. Malayalam is a free-word order language and words are seen agglutinated, hence most of the connectives are seen in agglutinated form. The discourse relation in Malayalam language can be syntactic (a suffix) or lexical[8]. Discourse relation can be within a clause, inter-clausal or inter-sentential. Discourse connectives are an important part of modeling discourse structure. In this paper, we now describe various connectives and arguments present in Malayalam language and the sense of relations. The two major categories of relations are Explicit and Implicit relations. We also observed the other types of relations such as Entity relation, Alternative lexicalized relation and No relation between the arguments.

3.1 Explicit Connectives

The explicit connectives are morphemes or free words that trigger discourse relations in Malayalam language. The connectives can occur at the initial, medial or final position of arguments in Malayalam language[9]. Here shows the example for explicit connective in Malayalam language.

(2)[prakqthi sundarangngaL anya pradESAngngaL kErALaththil dhArALamuNTu.]

Natural beauty places kerala+in many.

AtheykOTu <CON>

therefore

<ARG2>[vinOdasaFcaArikaL kErALaththe ishTapetunnu.]</ARG2>

Tourist+people kerala like.

(There are many natural beauty places in kerala. Therefore many tourist people like to visit kerala.)

The connective “therefore” in the above example occurs inter-sententially by connecting the two sentences. Connective occur at the initial position in the second argument. We see that the connectives are explicitly realizing the relations between two arguments.
There are different types of explicit connectives. Subordinate Conjunctions are one of the important types of connectives which connect the main clause with the adverbial clause, adjectival clause or a noun. Subordinate conjunctions occur as intra sentential connective, connecting two clauses mainly the main clause with the adverbial clause. Most commonly observed subordinate conjunctions are since, because and when. Co-ordinate Conjunctions give equal emphasis for two clauses. They connect two words, phrases and clauses. The most commonly observed co-ordinate conjunction in the malayalam corpus are “but” and “and”. Conjunct Adverbs join independent clauses together. These are special type of conjunctions as they are part of adverbs and conjunction. Correlative conjunction are another type of simple pair of conjunctions used in a sentence to join group of words or different words in text. This conjunction is not used to connect sentences themselves. But they link two or more words or clauses of equal importance within a sentence itself. Complementizer clause is considered as a special type of connective. It is a type of conjunction which marks a complement clause.

3.2. Implicit Connectives

An implicit relation can be inferred if there exist a relationship between adjacent pair of sentences and explicit connective is not present in the text. We have labeled as “IMPLICIT” label where an implicit relation was inferred[6].

(3) [Raamu skoolil poyi] <IMPLICIT> (pakshe) [Teechar klaasil
Ramu school went (but) Teacher class
didn't come

(Ramu went to school <IMPLICIT>(but)Teacher didn't come to class)

In the above example, two sentences are not explicitly connected but a relationship can be inferred implicitly.

3.3. AltLex, EntRel and NoRel

According to PDTB, there are Three types of relations AltLex, EntRel and NoRel can be annotated along with Explicit and implicit connectives[3]. We also observed these relations while annotating the text in malayalam language. Entity relation(EntRel) is one in which no discourse relation can be inferred and one argument describes entity in the other argument-tagged as <EntRel>. One argument alternatively lexicalized by other argument-tagged as <AltLex>. When none of the discourse relation between two adjacent sentences-tagged as <NoRel>.

4. Malayalam text corpus and annotation Process

We collected Malayalam corpus of 6000 sentences from kerala tourism website and annotated the connectives and arguments with effective and encouraging results. In Malayalam discourse, connectives can occur within a sentence or between sentences. In Malayalam inter sentence connectives are said to occupy sentence initial position. In our annotation, the first argument as <ARG1> and </ARG1>, the second argument as <ARG2> and </ARG2> and the connective as <CON>. Our annotation process also captures the various types of discourse connectives and their senses based on Penn Tree Discourse Tree Bank guidelines. The process of discourse annotation involves identifying discourse connectives in the raw text and annotating their arguments with semantics. We have annotated explicit connectives, implicit connectives and other types of connectives such as Altlex, EntRel and NoRel which shows the position of discourse relation in Malayalam language. We have observed the following points while annotating the text. 1) Identify the clause or phrases of Malayalam discourse 2) Determine how the relation exist in between the arguments based on the usage of text. 3) Characterize the type of relation in between the arguments and identify the sense of relation

5. Sense Annotations of Malayalam Connectives

Malayalam is an agglutinated language and most of the connectives are appeared in agglutinated form. Discourse connectives in Malayalam language are important signals for discourse relation recognition. Connectives can be used to relate the use of the arguments of a connective in text to one another or the use of one argument with the sense of the other argument. According to Penn Discourse Tree Bank [1], Explicit discourse connective in text has four senses, i.e., Temporal, Contingency, Comparison and Expansion. Temporal Conjunctions are identified when the situations or events expressed in the arguments of the relations are related temporally. The connectives when, while, before and after have a temporal sense. In contingency
class, conditional sense is identified when argument2 sets up a truth condition for argument1. Comparative sense is identified when two terms of the arguments are compared. Expansion sense is identified when argument1 extends its relation to argument2. We used various mnemonics while tagging the senses of relations. Some examples are given below.

(4) [ara glAs veLLam kuTichchathinu SEsham<CON><TEMP-ASYNC>]</ARG1>
half glass water drink after
[yoga pariSIlanam ArambhikkAm] </ARG2>
yoga practice will+start

(You can start practising yoga after drinking half glass of water)

(5) [chilayALukaL avarute SarIravEdana mARAn vEdana samhArikaL]
Some+people their body+pain relief pain killer
upayOgikkunnuNTu] </ARG1>, ennAl <COMP-OPPO> [ athu guNaththEkkAL used
However that quality+than
dOsham CeVyyum. ]</ARG2>
harmful will+do

(Some people use painkiller for getting relief from body pain. However it is more harmful to our body than its qualities.)

(6) [AvaSyaththinu uRakkam kiiTTAathevarumpOL <CONT-REAS>] </ARG1>
Enough sleep not+getting +when+come
[SarIraththinu AyAsam body pain
anubhavappetunnu. ]</ARG2>
suffering

(When we are not getting enough sleep, we get body pain.)

(7) [avar kqshikkuLL sthalam kaNTupitikkuka yum <EXP-CONJ>]</ARG1>
They farm place found + and
[viththum vaLavum vitharaNm ceVyyukayum uNTAyi] </ARG2>
Seed+and fertilizers distributing done happened

(They found out the place for farming and distributed the seed and Fertilizers.)

6. Results and Discussions

In our experiment, we have done annotations for 6000 sentences from Kerala tourism corpus and the statistics of connectives are obtained. The two arguments of the relation are being labeled as arg1 and arg2. In our approach the label assignment is syntactic. Sometimes, the arguments can be in the same sentence as the connective. Sometimes, one of the preceding sentence acts as an argument and the argument can be a non-adjacent sentence. But the text span follows the minimality-principle. Those connectives that do not occur as free words were considered to be part of arg1 and the other relation will be arg2. Here we present some useful distributions of various types of connectives identified in Malayalam language and its arguments. Annotated Malayalam corpus consists of 1744 connectives which includes explicit, implicit, alternative lexical(AltLex), Entity relation(EntRel) and no relation between arguments(NoRel). Connective types, connective total and count% are described in Table 1. While annotating explicit and implicit connectives, inter clausal and inter sentential rules are applied in the text based on the semantic properties of discourse as given in examples (4), (5), (6) and (7).

Table 2 shows the twelve relation types for modeling the text. Four major semantic categories Temporal, Contingency, Comparison and Expansion are highest level of relations. Most temporal relations are marked with connective like ‘before’, ‘after’ and ‘when’. The first and last word of the sentences turned out to be useful indicators of temporal relations. We identified different senses of ‘when’ connective as conditional sense, purely temporal sense and simultaneously temporal and causal sense. Contingency relation occurs in different sense labels which denotes the cause and effect of the arguments of relation. When cause and result of the situation occurs in the arguments of relation, we identified the sense class as ‘Contingency cause result’.
In expansion class, conjunction relation occurs between the events that is suffixed to the verb that describes each event. Some connectives have multiple senses based on the events or situation expressed in the arguments of the connective. In comparison, opposite sense is identified when different values are assigned to the terms of the arguments that are compared.

Table 1. Annotation statistics of connectives (Tagging Malayalam Corpus)

<table>
<thead>
<tr>
<th>Malayalam Corpus</th>
<th>Connective</th>
<th>Connective Count</th>
<th>Connective Count %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connective Types</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explicit</td>
<td>1509</td>
<td></td>
<td>86.53</td>
</tr>
<tr>
<td>Implicit</td>
<td>40</td>
<td></td>
<td>2.29</td>
</tr>
<tr>
<td>EntRel</td>
<td>123</td>
<td></td>
<td>7.05</td>
</tr>
<tr>
<td>Altlex</td>
<td>47</td>
<td></td>
<td>2.69</td>
</tr>
<tr>
<td>NoRel</td>
<td>25</td>
<td></td>
<td>1.43</td>
</tr>
</tbody>
</table>

Table 2. Sense annotation statistics of Malayalam Discourse

<table>
<thead>
<tr>
<th>Sense Annotations Class</th>
<th>Relation Types</th>
<th>Sense-Mnemonic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEMPORAL</td>
<td>Temporal Asynchronous</td>
<td>TEMP-ASYNC</td>
<td>161</td>
</tr>
<tr>
<td></td>
<td>Temporal Synchronous</td>
<td>TEMP-SYNC</td>
<td>56</td>
</tr>
<tr>
<td>CONTIGENCY</td>
<td>Contingency Conditional</td>
<td>CONT-COND</td>
<td>362</td>
</tr>
<tr>
<td></td>
<td>Contingency Cause Reason</td>
<td>CONT_CAUS_REAS</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Contingency Reason</td>
<td>CONT-REAS</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td>Contingency Cause Result</td>
<td>CONT_CAUS_RESU</td>
<td>200</td>
</tr>
<tr>
<td>COMPARISON</td>
<td>Comparison opposition</td>
<td>COMP-OPPO</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Comparison Concession</td>
<td>COMP_CONCESS</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>Comparison Contrast</td>
<td>COMP-CONTR</td>
<td>50</td>
</tr>
<tr>
<td>EXPANSION</td>
<td>Expansion Conjunction</td>
<td>EXP-CONJ</td>
<td>291</td>
</tr>
<tr>
<td></td>
<td>Expansion Alternative Conjunction</td>
<td>EXP-ALTER</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Expansion Restatement</td>
<td>EXP-REST</td>
<td>21</td>
</tr>
</tbody>
</table>
The lowest level senses for each sense annotation class give more meaningful model of text while annotating the Malayalam discourse. Each lower level relation types have assigned mnemonic for tagging sense annotations of Malayalam text.

7. Conclusion

We have reported the annotation results of Malayalam corpus which gives the clear structure of discourse connectives and their arguments in Malayalam language that can be reliably annotated for large corpus. We have presented our work on tagging of connectives and arguments for both explicit and implicit connectives in Malayalam language. We also reported the sense annotation of connectives while tagging the Malayalam discourse. We developed an annotated Malayalam discourse relation which helps for further research in Malayalam discourse. In future, we extend our work involving multiple annotators which gives more information of discourse connectives with inter-annotator agreement.

References