Creating an Annotated Tamil Corpus as a Discourse Resource

Ravi Teja Rachakonda
International Institute of
Information Technology
Hyderabad, India
raviteja.r@research.iiit.ac.in

Dipti Misra Sharma
International Institute of
Information Technology
Hyderabad, India
dipti@iiit.ac.in

Abstract

We describe our efforts to apply the Penn Discourse Treebank guidelines on a Tamil corpus to create an annotated corpus of discourse relations in Tamil. After conducting a preliminary exploratory study on Tamil discourse connectives, we show our observations and results of a pilot experiment that we conducted by annotating a small portion of our corpus. Our ultimate goal is to develop a Tamil Discourse Relation Bank that will be useful as a resource for further research in Tamil discourse. Furthermore, a study of the behavior of discourse connectives in Tamil will also help in furthering the cross-linguistic understanding of discourse connectives.

1 Introduction

The study of discourse structure in natural language processing has its applications in emerging fields such as coherence evaluation, question answering, natural language generation and textual summarization. Such a study is possible in a given human language only if there are sufficient discourse annotated resources available for that language. The Penn Discourse Treebank (PDTB) is a project whose goal is to annotate the discourse relations holding between events described in a text. The PDTB is a lexically grounded approach where discourse relations are anchored in lexical items wherever they are explicitly realized in the text (Miltsakaki et al. 2004, Prasad et al., 2008). To foster cross-linguistic studies in discourse relations, projects similar to the PDTB in discourse annotation were initiated in Czech (Mladová et al., 2008), Chinese (Xue, 2005), Turkish (Zeyrek and Webber, 2008) and Hindi (Prasad et al., 2008). We explore how the underlying framework and annotation guidelines apply to Tamil, a morphologically rich, agglutinative, free word order language.

In this paper, we present how a corpus of Tamil texts was created on which we performed our pilot experiment. Next, in Section 3 we cover the basics of the PDTB guidelines that we followed during our annotation process. In Section 4, we show various categories of Tamil discourse connectives that we identified after a preliminary study on discourse connectives in Tamil, illustrating each with examples. In Section 5, we discuss some interesting issues specific to Tamil that we encountered during discourse annotation and present the results of the pilot experiment that we performed on our source corpus. We conclude this paper in Section 6 by discussing about challenges that were unique to our work and our plans for the future.

2 Source Corpus

We collected Tamil encyclopedia articles from the June 2008 edition of the Wikipedia static HTML dumps[1]. Elements such as HTML metadata, navigational links, etc. were then removed until only the text of the articles remained. A corpus was then built by collecting the texts from all the articles in the dump. The corpus thus created consists of

[1] http://static.wikipedia.org/

Proceedings of the Fifth Law Workshop (LAW V), pages 119–123,
Portland, Oregon, 23-24 June 2011.

about 2.2 million words from approximately 200,000 sentences.

Since the texts used in building the corpus were all encyclopedia articles featured in the Tamil language version of Wikipedia, the corpus covers a wide variety of topics including arts, culture, biographies, geography, society, history, etc., written and edited by volunteers from around the world.

3 Penn Discourse Treebank Guidelines

The PDTB is a resource built on discourse structure in (Webber and Joshi, 1998) where discourse connectives are treated as discourse-level predicates that always take exactly two *abstract objects* such as events, states and propositions as their arguments. We now describe the types of connectives and their senses from the PDTB framework and provide examples from Tamil sentences.

3.1 Annotation Process

The process of discourse annotation involves identifying discourse connectives in raw text and then annotating their arguments and semantics. Discourse connectives are identified as being *explicit, implicit, AltLex, EntRel* or *NoRel* (Prasad et al. 2008). These classes are described in detail in Section 4. By convention, annotated *explicit* connectives are underlined and *implicit* connectives are shown by the marker, "(Implicit=)". As can be seen in example (1), one of the arguments is shown enclosed between {} and the other argument is shown in []. The *AltLex, EntRel* or *NoRel* relations are shown by underlining, i.e., as "(AltLex=)", "(EntRel)" and "(NoRel)", respectively.

(1) {eN kAl uDaindadaN}Al [eNNAl viLayADa muDiyAdu].
'{My leg broke}, hence [I cannot play].'

3.2 Sense Hierarchy

The semantics of discourse relations are termed as *senses* and are then classified hierarchically using four top-level *classes* 'Comparison', 'Contingency', 'Expansion' and 'Temporal'. Each class is refined by its component *types* and these, in turn, are further refined by the *subtype* level.

It is interesting to note that some connectives have multiple senses. In example (2) the affixed *–um* connective carries the sense of type *Expansion:Conjunction* 'also' whereas in example (3) the same affix carries the sense of the subtype *Contingency:Concession* 'however'.

(2) {idaN mUlam avar oru nAL pOttiyil oNbadAyiram OttangaLai kaDanda pattAvadu vIrar eNra perumaiyai pettrAr}. [inda OttangaLai kaDanda mudal teNNAppirikka vIrar eNra sAdaNaiyaiy]um [nigaztiNAr].
'{By this, he became the tenth player to cross nine thousand runs in one-day internationals}. [He] also [attained the record of becoming the first South African player to cross these many runs].'

(3) {seNra murai kirikket ulagakkOppaiyiN pOthu pangu pattriyadai vida iraNDu aNigaL immurai kUDudalAga pangu pattriya pOd}um, [motthap pOttigaL inda muraiyil kuraivANadAgum.]
'Though {two more teams participated when compared to last Cricket World Cup}, [the total matches played during this time were fewer].'

4 Discourse Connectives in Tamil

Tamil is an *agglutinative language* where morphemes are affixed to the roots of individual words, a trait that it shares with many other Dravidian languages and languages like Turkish, Estonian and Japanese. Here, each affix represents information such as *discourse connective, gender, number*, etc. We now describe how we try to capture various types of Tamil discourse connectives using a proposed scheme which is based on the existing PDTB guidelines proposed by (Prasad et al., 2007).

4.1 Explicit Discourse Connectives

Explicit discourse connectives are lexical items present in text that are used to anchor the discourse relations portrayed by them. In Tamil, they are found as affixes to the verb, as in example (4) where the affix *-Al* conveys the meaning 'so'. This is in a way similar to the *simplex subordinators* in Turkish, as described in (Zeyrek and Webber, 2008). However, like in English, explicit discourse connectives are also realized as unbound lexical items, as can be seen in example (5) where the word *eNavE* means 'hence'.

(4) {avaradu uDalnalam sariyillAmaiy}Al [nAngu mAdangaL avarAl viLayADa iyalavillai].
'{He was suffering from ill health} so [he could not play for four months].'

(5) {tirukkuraL aNaittu madattiNarum paDittu payaNaDaiyum vagaiyil ezudappattuLLadu}. eNavE [innUl palarAl pArAttappaDuginradu].
'{Thirukkural has been written in such a way that people from all religions can benefit from it}. Hence, [this book is praised by many].'

Syntactically, explicit connectives can be *coordinating conjunctions* e.g., *alladu* ('or'), *subordinating conjunctions* e.g., *-Al* ('so'), *sentential relatives* e.g., *-adaNAl* ('because of which'), *particles* e.g., *-um* ('also') or *adverbials* e.g., *-pOdu* ('just then').

Explicit connectives also occur as *conjoined connectives* where two or more instances of connectives share the same two arguments. Such connectives are annotated as distinct types and are annotated discontinuously, as seen in example (6) where the connectives *-um* and *-um* are paired together to share the same arguments.

(6) {mANavargaLukku sattuNavu aLikkav}um [avargaL sariyAga uDarpayirchi seiyyav]um arasup paLLigaL udava vENDum.
'Government schools should help in {providing nutritious food to the students} and [making sure they perform physical exercises].

4.2 Implicit Discourse Connectives

Implicit discourse connectives are inserted between adjacent sentence pairs that are not related explicitly by any of the syntactically defined set of explicit connectives. In such a case, we attempted to infer a discourse relation between the sentences and a connective expression that best conveys the inferred relation is inserted. In example (7), the implicit expression *uthAraNamAga* ('for example') has been inserted as an inferred discourse relation between the two sentences.

(7) {IyOrA iNa makkaLiN moziyil irundu iNru Angilattil vazangum sorkaL uLLaNa}. (Implicit=uthAraNamAga) [dingO, vUmErA, vAlabi pONra sorkaL IyOravilirindu tONriya sorkaL dAN].
'{There are words that are present in English that originated from the language of the Eora people}. (Implicit= For example) [Dingo, Woomera and Wallaby are words with their origins in Eora].'

4.3 AltLex, EntRel and NoRel

In cases where no implicit connective was appropriately found to be placed between adjacent sentence-pairs, we now look at three distinct classes. *AltLex* relations, as seen in example (8) are discourse relations where the insertion of an implicit connective leads to a redundancy in its expression as the relation is already alternatively lexicalized by some other expression that cannot be labeled as an explicit connective. Example (9) shows an *EntRel* relation where no discourse relation can be inferred and the second sentence provides further description of an entity realized in the first sentence. When neither a discourse relation nor entity-based coherence can be inferred between the two adjacent sentences, it is described as a *NoRel*, shown in example (10).

(8) {mudalAvadAga mAgim, jOgEshwari, pUrivilla rayil nilayangaLil guNDu vedittadu}. (AtlLex=idai toDarndu) [mErku rayilvEyiN aNaittu rayilgaLum niruttappaTTaNa].
'{Initially, bombs exploded in Mahim, Jogeshwari and Poorivilla}. (AltLex=following this) [all the trains from the western railway were halted].'

(9) {ivvANDu kirikket ulagakkOppai mErkindiyat tIvugaLil mArc padimUnril irundu Epral irubattu-ettu varai naDaipettradu}. (EntRel) [indap pOttiyil pangupattriya padiNAru nADugaLaic cArnda aNigaLum ovvoru kuzuvilum nANgu aNigaL vIdamAga nANgu kuzukkaLAga pirikkapattu pOttigaL iDampettraNa].
'{This year's Cricket World Cup was held in West Indies from the thirteenth of March to the twenty-eight of April}. (EntRel) [In this competition, the teams representing the sixteen nations were grouped into four groups with four teams in each group].'

(10) {caccin TeNdUlkar ulagiNilEyE migac ciranda mattai vIccALarAga karudappadugirAr}. (NoRel) [indiya pandu vIccALargaL sariyANa muraiyil payirci peruvadillai].
'{Sachin Tendulkar is considered the best batsman in the world}. (NoRel) [Indian bowlers are not being given proper coaching].'

5 Observations and Results

5.1 Combined connectives

There is a paired connective *-um ... -um (...)* that sometimes expresses an *Expansion:Conjunction* relation between the events where each *-um* is suffixed to the verb that describes each event (see example (6)). Also, there is a connective *-Al* which usually never occurs more than once and sometimes expresses a *Contingency:Cause* relation between two events.

It is interesting to see that in sentences like (11), the *-Al* combines with the *-um ... -um* to express something like a new type of relation. In the process, the *-um ... -um* causes the *-Al*, which is usually not doubled, to become doubled, thereby forming an *-Alum ... -Alum*. We call this special type of connectives as *combined connectives*, as shown in example (11).

(11) {kirikket viLayADiyad}Alum {uDarpayirci seidad}Alum [sOrvaDaindEN].
'Because {I played cricket} and because {I did exercise} [I am tired].'

5.2 Redundant connectives

The connective *-O ... -O (...)* that conveys a *dubitative* relation also combines with the *-Al* connective in a way similar to what was shown in Section 5.1 to form the combined connective *-AlO ... -AlO (...)*.

However, in example (12), *alladu*, an equivalent of the *-O ... -O* connective has also occurred in addition to the combined *-AlO ... -AlO* connective. This may be purely redundant, or could serve a purpose to emphasize the dubitative relation expressed by both *alladu* and *-O ... -O*.

(12) {pOtti samappatt}AlO alladu {muDivu perapaDAmal pON}AlO [piNvarum muraigaL mUlam aNigaL tarappaDuttapaDum].
'If {a game is tied} or if {there is no result}, [the qualified teams are chosen using the following rules].'

5.3 Results of Pilot Study

In this experiment, we looked at 511 sentences from the corpus mentioned in Section 2 and annotated a total of 323 connectives. Table 1 shows the distribution of the annotated connectives across the different types such as Explicit, Implicit, EntRel, AltLex and NoRel.

Connective Type	Count	Count (unique)	Count (%)	Senses
Explicit	269	96	83.3	18
Implicit	28	16	8.6	13
EntRel	16	-	5.0	-
AltLex	8	5	2.5	4
NoRel	2	-	0.6	-

Table 1: Results of Pilot Experiment

While a higher percentage of the connectives annotated are those of the Explicit type, it can also be seen that there is a higher proportion of unique connectives in the Implicit and AltLex types. Note that since EntRel and NoRel connectives are not associated with a sense relation or a lexical item, their counts are left blank.

6 Challenges and Future Work

The agglutinative nature of the Tamil language required a deeper analysis to look into suffixes that act as discourse connectives in addition to those that occur as unbounded lexical items. We also found certain interesting examples that were distinct from those observed during similar approaches in relatively less morphologically rich languages like English.

While this was a first attempt at creating a discourse annotated Tamil corpus, we are planning to conduct future work involving multiple annotators which would yield information on annotation metrics like inter-annotator agreement, for example. Our work and results would also be useful for similar approaches in other morphologically rich and related South Indian languages such as Malayalam, Kannada, Telugu, etc.

We will also work on a way in which the discourse annotations have been performed will help in augmenting the information provided during dependency annotations at the sentence-level.

Acknowledgments

We are grateful to Prof. Aravind Joshi and Prof. Rashmi Prasad of University of Pennsylvania and Prof. Bonnie Webber of University of Edinburgh for their valuable assistance and feedback.

We would like to thank Prof. Rajeev Sangal of IIIT Hyderabad for his timely guidance and useful inputs. We also acknowledge the role of Sudheer Kolachina in the discussions we had in the writing of this paper.

References

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber. 2004. The Penn Discourse Treebank. Proceedings of LREC-2004.

Rashmi Prasad, Samar Husain, Dipti Mishra Sharma and Aravind Joshi. 2008. Towards an Annotated Corpus of Discourse Relations in Hindi. Proceedings of IJCNLP-2008.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo and Bonnie Webber. 2007. The Penn Discourse Tree Bank 2.0 Annotation Manual. December 17, 2007.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, Discourse Relations and Discourse Markers: Proceedings of the Conference, pages 86-92. Association of Computational Linguistics.

Nianwen Xue. 2005. Annotating Discourse Connectives in the Chinese Treebank. Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky.

Deniz Zeyrek and Bonnie Webber. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. Proceedings of IJCNLP-2008.