Cross Linguistic Variations in Discourse Relations among Indian Languages

Sindhuja Gopalan
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
sindhujagopalan@au-kbc.org

Lakshmi s
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
lakssreedhar@gmail.com

Sobha Lalitha Devi
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Abstract

This paper summarizes our work on analysis of cross linguistic variations in discourse relations for Indo-Aryan language Hindi and Dravidian languages Malayalam and Tamil. In this paper we have also presented an automatic discourse relation identifier, which gave encouraging results. Analysis of the results showed that some complex structural inter-dependencies existed in these three languages. We have described in detail the structural inter-dependencies that occurred. Discourse relations in the three languages thus exhibited complex nature due to the structural inter-dependencies.

1 Introduction

Discourse relations link clauses in text and compose overall text structure. Discourse relations are used in natural language processing (NLP), including text summarization and natural language generation. The analysis and modeling of discourse structure has been an important area of linguistic research and it is necessary for building efficient NLP applications. Hence the automatic detection of discourse relation is also important. The Indo-Aryan (Hindi) and Dravidian languages (Malayalam and Tamil) share certain similarities such as verb final language, free word order and morphologically rich inflections. Due to the influence of Sanskrit in these languages they are similar at lexical level. But structurally they are very different. In this work we have presented an analysis of the cross linguistic variations in the discourse relations among three languages Hindi, Malayalam and Tamil. Instead of identifying all possible discourse relations we have considered the analysis of explicit discourse relations and developed an automatic discourse relation identification system. During error analysis various structural interdependencies were also noted.

Discourse tagging for Indian languages Hindi, Malayalam and Tamil has been done by Sobha et al., (2014) Other published works on discourse relation annotations in Indian languages are in Hindi (Kolachina et al., (2012); Oza et al., (2009)) and Tamil (Rachakonda and Sharma (2011)). Menaka et al., (2011) in their paper have automatically identified the causal relations and have described about the structural interdependencies that exist between the relations. Similarly, we observed the existence of structural interdependencies between the discourse relations in three languages, which we have explained in detail. From the previous works on discourse relation annotation for various Indian languages, we can observe that the study of discourse relations is carried out for specific Indian language and hence we attempted to discuss the cross linguistic variations among Hindi, Tamil and Malayalam languages.

Researchers have performed identification and extraction of discourse relation using cue based or statistical methods. Penn Discourse Tree Bank (PDTB) is the large scale annotated corpora of linguistic phenomena in English (Prasad et al., 2008). The PDTB is the first to follow the lexically grounded approach to annotation of discourse relations. Marcu and Echihabi (2012) have focused on recognition of discourse relation using cue phrases, but not extraction of arguments. Wellner and Pustejovksy (2007) in their study considered the problem of automatically identifying the arguments of discourse connectives in PDTB. They re-casted the problem to that of identifying the argument heads, instead of

Proc. of ICON-2017, Kolkata, India. December 2017 ©*2017 NLPAI*, pages 402–407

identifying the full extents of the arguments as annotated in PDTB. To address the problem of identifying the arguments of discourse connectives they incorporated a variety of lexical and syntactic features in a discrimination log-linear re-ranking model to select the best argument pair from a set of N best argument pairs provided by independent argument models. They obtained 74.2% accuracy using gold standard parser and 64.6% accuracy using automatic parser for both arguments. Elwell and Baldridge (2008) have used models tuned to specific connectives and connective types. Their study showed that using models for specific connectives and types of connectives and interpolating them with a general model improves the performance. The features used to improve performance include the morphological properties of connectives and their arguments, additional syntactic configuration and wider context of preceding and following connectives. The system was developed on PDTB. They used Maximum entropy ranker. Models were trained for arg1 and arg2 selection separately. They achieved 77.8% accuracy for identifying both arguments of connective for gold standard parser and 73.6% accuracy using automatic parser. Ramesh and Yu (2010) have developed a system for identification of discourse connectives in bio-medical domain. They developed the system on BioDRB corpus using CRFs algorithm. For PDTB data they obtained F-score of 84%. They obtained F-score of 69% for BioDRB data. For PDTB based classifier on BioDRB data, they obtained F-score of 55%. In this work they did not focus on identification of arguments. Versley (2010) presented his work on tagging German discourse connectives using a German–English parallel corpus. AlSaif (2012) used machine learning algorithms for automatically identifying explicit discourse connectives and its relations in Arabic language. Wang et al., (2012) used sub-trees as features and identified explicit and implicit connectives and their arguments. Zhou et al., (2012) presented the first effort towards cross lingual identification of the ambiguities of discourse connectives. Faiz et al., (2013) did explicit discourse connectives identification in the PDTB and the Biomedical Discourse Relation Bank (BDRB) by combining certain aspects of the surface level and syntactic feature sets. In this study we tried to develop a discourse parser for all three languages for identification of connectives and its arguments.

Following sections are organized as follows. Corpus Collection and Annotation is described in section 2, cross linguistic variations in discourse relations among three languages is given in section 3, method used for the automatic identification of discourse relation and the results are described in section 4 and the various structural interdependencies that occur in the three languages is described in section 5. The paper ends with the conclusion section.

2 Corpus collection and Annotation

Health related articles were chosen from web and after removing inconsistencies like hyperlinks a total corpus of 5000 sentences were obtained. Then we annotated the corpus for connectives and its arguments. The discourse relation annotation was purely syntactic. The arguments were labeled as arg1 and arg2 and arg2 was chosen to be following arg1. When free words occur, we tag them separately and the discourse unit between which the relation is inferred is marked as arg1 and arg2. When the connectives exist as bound morphemes we keep them along with the word to which it is attached and include it under arg1. The annotated corpus contains 1332 explicit connectives in Hindi, 1853 in Malayalam and 1341 in Tamil. From the data statistics we can observe that Malayalam language has more number of connectives than Tamil and Hindi. Annotated corpus is used to train the system and the models are built for the identification of connectives and arguments.

3 Cross Linguistic variations in Discourse Relations

The discourse relation in Indian language can be expressed in many ways. It can be syntactic (a suffix) or lexical. It can be within a clause, inter-clausal or inter-sentential. The various cross linguistic variations in discourse relation among the three languages is analyzed and described below.

3.1 Discourse Connectives

Discourse relations can be inferred using Explicit or Implicit connectives. Explicit connectives connect two discourse units and trigger discourse relation. The explicit connectives can be realized in any of the following ways.

- Subordinators that connect the main clause with the subordinate or dependent clause. (For example: agar-to, jabkI in Hindi, appoL, -aal in Malayalam and -aal, ataal in Tamil).
- Coordinators which connect two or more items of equal syntactic importance.

They connect two independent clauses. (For example: “aur”, “lekin” in Hindi, “-um”, “ennaal” in Malayalam and “anaal”, “athanaal” in Tamil).

- Conjunct adverbs that connect two independent clauses and modify the clauses or sentences in which they occur. (For example: “isliye”, “halaanki” in Hindi, “athinaal”, “aakayaal” in Malayalam and “enninum”, “aakaiyaal” in Tamil).
- Correlative conjunctions which are paired conjunctions. They link words or group of words of equal weights in a sentence. (For example: “na keval balki” in Hindi, “maathramalla-pakshe” in Malayalam and “mattumalla-aanaal” in Tamil).

3.2 Position of Connectives

In our approach we have done a syntactic based tagging. In Hindi, Malayalam and Tamil discourse connectives can occur within a sentence or between sentences. In all the three languages inter sentence connectives are said to occupy sentence initial position. Example 1 shows the inter sentence discourse relation in Malayalam.

Example 1:

[chila aaLukaL mukhsoundaryam koottaan
Some people facial-beauty increase
kreemukaL upayogikkaaruNt.]/arg1
creams use
ennaal [athu guNathekkaaLeRe doshamaaN
But that goodness-more than harm-is
cheyyuka.]/arg2
do

(Some people use creams to increase their facial beauty. But that will do more harm than good.)

We found that there exists a difference in the position of conjunct adverb “although” among the three languages. As in Example 2, in Hindi this connective occurs in the sentence initial position whereas in Tamil and Malayalam this connective occurs in the middle position and remains agglutinated with the verb.

Example 2:

haalaaMki [yoga pakshaaGaath kii samasyaa kaa
although yoga paralysis problem's
sTaayii samaaDhaan karthaa hai]/arg2,
permanent solution do is
[yah samay lethaa hai evaM shramsaaDya
This time take is and painstaking
hai]/arg1
is

(Although yoga gives a permanent solution for paralysis, this is time taking and painstaking.)

In Tamil and Malayalam the connective “and” exists in the form as in Example 3. In Hindi single lexicon “aur” serves this purpose.

Example 3:

[muuttukaLiluLLa kuRuththelumpu vaLaraamal
in knee cartilage without
theymaanam atainthaal**um**]/arg1,
growing wear if get-and
[angkuLLa vazhuvazhuppaana thiravam
there smooth fluid
kuRainthupoonaal**um**]/arg2 muuttukaLil uraayvu
get less-and knee friction
eRpatum.
will develop

(If cartilage in the knee gets wear without growing and if the smooth fluid present there becomes less, friction will develop in the knee.)

3.3 Agglutinated and intra sentence

In Malayalam and Tamil connectives can occur as free words or bound morphemes. But in Hindi only free word connectives exist as in Example 2.

Example 4:

[vayiRRil kutalpun irunth**aal**]/arg1 [vayiRu
In stomach ulcer is there-if stomach
valikkum]/arg2.
will pain

(If there is ulcer in stomach, stomach will pain.)

3.4 Paired connectives

In Hindi some discourse connectives were seen as paired connectives. This type of connectives is not noticed in Malayalam and Tamil.

Example 5:

yadhii [lagaathaar buKaar aa rahaa hai]/arg1 **tho**
if constantly fever coming is then
[uskii jaaNca avashaya karaaye]/arg2.
its check sure do

(If fever is coming constantly, then check it for sure.)

In the above Example 5 “yadhii-to” is the paired connective that occurs at the start of arg1 and arg2. Whereas in Tamil and Malayalam it occurs as a single connective as in Example 4 and occurs agglutinated with verb.

3.5 Arguments of Relations

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. Also the argument can be a non-adjacent sentence. But the text span follows the minimality-principle. In Example 1 the connective “ennal” in Malayalam

connects two discourse units inter sententially. The discourse unit that follows the connective is arg2 and the preceding unit is arg1. In Example 4 the arguments for connective "-aal" in Tamil occur in same sentence.

4 Automatic identification of discourse relation

4.1 Method Used

We have used the method adopted by Menaka et al., (2011) for the identification of discourse relations. We have preprocessed the text for morph analysis (Ram et al, 2010), part-of-speech tagging (PoS) (Sobha et al, 2016), chunking (Sobha and Ram, 2006), clause tagging (Ram et al, 2012). The implementation is done based on machine learning technique CRFs.

4.2 Conditional Random Fields

CRFs is an undirected graphical model, where the conditional probabilities of the output are maximized for a given input sequence. We chose CRFs, because it allows linguistic rules or conditions to be incorporated into machine learning algorithm. Here, we have used CRF++ (Kudo, 2005), an open source toolkit for linear chain CRFs.

4.3 Features Used

For the identification of connectives, we have used PoS tagging information, morphological suffixes and clause information as features for Malayalam and Tamil. Morphological suffixes such as conditional markers, causal markers, relative participle (RP) marker followed by postposition (PSP) and coordination markers were used. For connective identification in Hindi, word, PoS tagging information and chunk information were used. For argument identification we have taken PoS tagging information, chunk information, morphological suffixes, and clause information, combination of PoS and chunk information and connectives as features.

4.4 Training and Testing

For identifying the discourse connectives, we trained the system using the features for connectives. In the next stage we train the system to identify the arguments and their text spans. Here we have built 4 language models for each of the 4 boundaries – Arg2-START, Arg1-END, Arg1-START and Arg2-END motivated by the work of Menaka et al., (2011). The system was trained in 4 phases to develop 4 models. We used 4000 sentences from the corpus for training and 1000 sentences for testing. For testing, the sentences are pre-processed similarly as training data. The system identified the discourse markers in stage 1 and this output becomes input to stage 2. In both the stages we used CRFs as the machine learning algorithm.

The performance of our system is measured in terms of Precision, Recall and F score. Precision is the number of discourse relations correctly perceived by the system from the total number of discourse relations identified, Recall is the number of discourse relations correctly detected by the system by the total number of discourse relations contained in the input text and F-score is the harmonic mean of precision and recall.

The results for connective identification are tabulated in Table 1.

	Precision	Recall	F-score
Hindi	96.33	92.3	94.27
Malayalam	96.3	91.6	93.89
Tamil	95.35	94.18	94.76

Table 1: Results for Connective Identification

The argument identification results are given in Table 2, Table 3, Table 4 and Table 5.

	Precision	Recall	F-score
Hindi	76	72.2	74.05
Malayalam	78.5	72	75.1
Tamil	81.53	73.6	77.36

Table 2: Results for ARG1 Start

	Precision	Recall	F-score
Hindi	75.9	72.2	74
Malayalam	78.8	72	75.23
Tamil	82	72.6	77

Table 3: Results for ARG1 End

	Precision	Recall	F-score
Hindi	77.4	73.2	75.24
Malayalam	79.2	73	75.97
Tamil	81.5	72.6	76.79

Table 4: Results for ARG2 Start

	Precision	Recall	F-score
Hindi	76.3	71.2	73.66
Malayalam	78.7	72.4	75.42
Tamil	82	72.7	77

Table 5: Results for ARG2 End

During error analysis it is noted that a good number of errors are due to structural interdependencies between discourse relations. When there are such structures, there is a considerable overlap in the arguments of two discourse relations leading to the improper identification of boundaries by the system. These are discussed in detail in the next section.

5 Structural Interdependencies between discourse relations

Some very unique pattern of interdependencies was seen existing between discourse relations for Hindi, Malayalam and Tamil mainly due to the free word order nature of those languages. Given below are such patterns.

5.1 Embedding within itself

Due to the free word order nature of Indian languages this type of structure comes into being. Consider the Malayalam Example 6 given below.

Example 6:

[pala padhathikaLum [ee karaaR
many plans this contract
sambhavikkaathathinaal]/arg1 natakkaathe
not-happen-hence failed
poyi.]/arg2

(This contract didn't happen, hence many plans failed.)

Here arg1 and marker is seen embedded inside arg2.

5.2 Between Two Discourse Relations – Containment

One most frequently occurring structural dependency is that of embedding or containment of the whole of a discourse relation within one of the arguments of another discourse relation.

Example 7:

[lagbhag 25 se 50 prathishath roobelaa
approximately 25 from 50 percent rubella
saMkramaN kaa pathaa nahiM cal paathaa]/arg1$_i$
infection know not get
aur [**agar[** iske lakshaN paidhaa hothe
and if its symptoms develop
haiM]/arg1$_j$ **tho** [[ve bhahuth hii
is then they very
halke hothe haiN]/arg2$_i$]/arg2$_j$
light is

(Approximately 25 to 50 percent of rubella infection is not known and if its symptoms develop then they are very light.)

The Example 7 shows that the arguments of connective "agar-to" are contained within the arg2 of connective "aur".

5.3 Between two Discourse Relations – Complete Overlap/Shared Argument

An argument may be shared by two discourse relations in different ways.

Example 8:

naviina vaazhkkai muRaiyil vaakanagkalaip
modern life style vehicles
payanpatutthuv**athaal**]/arg1$_i$ [[nataippayiRci
use-because walking
enpathu kuRainthuvittathu]/arg2$_i$]/arg1$_j$.
is reduced
ithanaal [utalil cerum
Because of this in body accumulate
thevaiyaRRa kalorikaL cariyaaka
unwanted calories correctly
erikkappatuvathillai]/arg2$_j$.
not burnt

(Because of using vehicles in modern life style walking is reduced. Because of this, the unwanted calories accumulated in the body is not burnt.)

In Example 8 the arg2 of the first discourse relation is the shared argument for the second discourse relation.

5.4 Completely Independent Relations

Example 9:

[poshakaaharam nalki kuttiye
nourishing-food gave child
paripaalichu.]/arg1$_i$ **engilum** [kuttiyute
fostered But child's
arogyathil purogathiyilla.]/arg2$_i$ [atuthaghathathil
health-in no-progress next-stage-in
guLikakaL nalki.]/arg1$_j$ **engilum** [kuttiyte
vitamin tablets gave But child's
arogyam athe nilayil thutarnnu.]/arg2$_j$
health same condition-in continued.

(Nourishing food was given for the child. But the child's health had no progress. In the next stage gave vitamin tablets. But the child's condition remained the same.)

In Example 9 there are two adjacent discourse relations which are independent of each other.

6 Conclusion

We have presented our work on discourse relation identification for Hindi, Malayalam and Tamil. An analysis of the discourse relations among the three languages was performed and an automatic identification system for discourse relation was developed. By analyzing the results

structural dependencies were noted. By handling this issue the performance of the system can be improved which makes up our future work.

Reference

John L AlSaif. 2012. Human and automatic annotation of discourse relations for Arabic, Ph.D. thesis, University of Leeds.

Ben Wellner and James Pustejovsky. 2007. Automatically Identifying the Arguments of Discourse Connectives, *Proceedings of EMNLP-CoNLL*, Prague, 92-101.

Balaji P. Ramesh and Hong Yu. 2010. Identifying discourse connectives in biomedical text, *Proceedings of AMIA Annual Symposium*, Washington, DC 657-661.

Daniel Marcu and Abdessamad Echihabi. 2012. An unsupervised approach to recognizing discourse relations, *Proceedings of 40th Annual Meeting on Association for Computational Linguistics*, 368-375.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers, *Proceedings of International Conference on Semantic Computing*, Santa Clara, CA.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0, *Proceedings of Language Resources and Evaluation Conference*, Marrakech, Morocco.

Lanjun Zhou, Wei Gao, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2012. Cross-Lingual Identification of Ambiguous Discourse Connectives for Resource-Poor Language, *Proceedings of International Conference on Computational Linguistics*, Mumbai, India, 1409-1418.

Ram, RVS, Bakiyavathi, T, Sindhujagopalan, R, Amudha, K and Sobha, L. 2012. Tamil Clause Boundary Identification: Annotation and Evaluation, *Proceedings of 1st Workshop on Indian Language Data: Resources and Evaluation*, Istanbul.

Ravi T. Rachakonda, and Dipti M. Sharma. 2011. Creating an annotated Tamil corpus as a discourse resource, *Proceedings of 5th Linguistic Annotation Workshop*, Portland, Oregon, 119-123.

Sudheer Kolachina, Rashmi Prasad, Dipti M. Sharma, and Aravind Joshi. 2012. Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank, *Proceedings of Language Resources and Evaluation Conference*, Istanbul, Turkey, 823-828.

S. Menaka, Pattabhi R.K. Rao, and Sobha L. Devi. 2011. Automatic identification of cause-effect relations in tamil using CRFs, *Proceedings of Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, 6608:316-327.

Sayeed I. Faiz, and Robert E. Mercer. 2013. Identifying explicit discourse connectives in text, *Advances in Artificial Intelligence, Lecture Notes in Computer Science*, 7884:64-76.

Sobha, L and Vijay Sundar Ram, R. 2006. Noun Phrase Chunker for Tamil, *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages (MSPIL)*, IIT Mumbai, India, 194-198.

Sobha L. Devi, S. Lakshmi, and Sindhuja Gopalan. 2014. Discourse Tagging for Indian Languages, *Proceedings of Computational Linguistics and Intelligent Text Processing*, Berlin, Heidelberg, 469-480.

Sobha L Devi, Pattabhi RK Rao and Vijay Sundar Ram, R. 2016. *AUKBC Tamil Part-of-Speech Tagger (AUKBC-TamilPoSTagger 2016v1)*, web download, http://www.au-kbc.org/nlp/corpusrelease.html.

Taku Kudo. 2005. CRF++, an open source toolkit for CRF, http://crfpp.sourceforge.net.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti M. Sharma, and Aravind Joshi. 2009. The Hindi discourse relation bank, *Proceedings of Third Linguistic Annotation Workshop*, 158-161.

Vijay Sundar Ram, R, Menaka, S and Sobha Lalitha Devi. 2010. Tamil Morphological Analyser", in "Morphological Analysers and Generators, LDC-IL, Mysore, 1 –18.

Xun Wang, Suj Ian Li, Jiwei Li, and Wenj Le Li. 2012. Implicit Discourse Relation Recognition by Selecting Typical Training Examples, *Proceedings of International Conference on Computational Linguistics*, Mumbai, India, 2757-2772.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection, *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, 83-82.