The Hindi Discourse Relation Bank

Umangi Oza[*], Rashmi Prasad[†], Sudheer Kolachina[*], Dipti Misra Sharma[*] and Aravind Joshi[†]
*Language Technologies Research Centre
IIIT Hyderabad, Gachibowli, Hyderabad, Andhra Pradesh, India 500032
oza.umangi,sudheer.kpg08@gmail.com,dipti@iiit.ac.in

[†]Institute for Research in Cognitive Science/Computer and Information Science
3401 Walnut Street, Suite 400A
Philadelphia, PA USA 19104
rjprasad,joshi@seas.upenn.edu

Abstract

We describe the Hindi Discourse Relation Bank project, aimed at developing a large corpus annotated with discourse relations. We adopt the lexically grounded approach of the Penn Discourse Treebank, and describe our classification of Hindi discourse connectives, our modifications to the sense classification of discourse relations, and some cross-linguistic comparisons based on some initial annotations carried out so far.

1 Introduction

To enable NLP research and applications beyond the sentence-level, corpora annotated with discourse level information have been developed. The recently developed Penn Discourse Treebank (PDTB) (Prasad et al., 2008), for example, provides annotations of discourse relations (e.g., causal, contrastive, temporal, and elaboration relations) in the Penn Treebank Corpus. Recent interest in cross-linguistic studies of discourse relations has led to the initiation of similar discourse annotation projects in other languages as well, such as Chinese (Xue, 2005), Czech (Mladová et al., 2008), and Turkish (Deniz and Webber, 2008). In this paper, we describe our ongoing work on the creation of a Hindi Discourse Relation Bank (HDRB), broadly following the approach of the PDTB.[1] The size of the HDRB corpus is 200K words and it is drawn from a 400K word corpus on which Hindi syntactic dependency annotation is being independently conducted (Begum et al., 2008). Source corpus texts are taken from the Hindi newspaper *Amar Ujala*, and comprise news articles from several domains, such as politics, sports, films, etc. We present our characterization of discourse connectives and their arguments in Hindi (Section 2), our proposals for modifying the sense classification scheme (Section 3), and present some cross-linguistics comparisons based on annotations done so far (Section 4). Section 5 concludes with a summary and future work.

2 Discourse Relations and Arguments

Following the PDTB approach, we take discourse relations to be realized in one of three ways: (a) as *explicit connectives*, which are “closed class” expressions drawn from well-defined grammatical classes; (b) as *alternative lexicalizations* (AltLex), which are non-connective expressions that cannot be defined as explicit connectives; and (c) as *implicit connectives*, which are implicit discourse relations “inferred” between adjacent sentences not related by an explicit connective. When no discourse relation can be inferred between adjacent sentences, either an *entity-based coherence relation* (called EntRel) or the absence of a relation (called NoRel) is marked between the sentences. The two abstract object relata of a discourse relation are called the relation’s arguments (named Arg1 and Arg2), and argument annotation follows the “minimality principle” in that only as much is selected as the argument text span as is minimally necessary to interpret the relation. Finally, each discourse relation is assigned a sense label based on a hierarchical sense classification.

2.1 Explicit Connectives

In addition to the three major grammatical classes of Explicit connectives in the PDTB – subordinating conjunctions, coordinating conjunctions, and adverbials – we recognize three other classes, described below.

[1] An earlier study of Hindi discourse connectives towards the creation of HDRB is presented in Prasad et al. (2008).

Sentential Relatives: These are relative pronouns that conjoin a relative clause with its matrix clause. As the name suggests, only relatives that modify verb phrases are treated as discourse connectives, and not those that modify noun phrases. Some examples are जिससे (so that), जिसके कारण (because of which).

1) [सारा काम छोड़कर वह उस चिड़िया को उठाकर दवा घर की ओर भागा] **जिससे** {उसका सही इलाज किया जा सके}

“[Dropping all his work, he picked up the bird and ran towards the dispensary], **so that** {it could be given proper treatment}.”

Subordinators: These include postpositions (Ex. 2), verbal participles, and suffixes that introduce non-finite clauses with an abstract object interpretation.[2]

2) [बा की बातें सुन]**कर** {मन-ही-मन गांधीजी बहुत लज्जित हुए}।

“**Upon** [hearing *Baa*’s words], {*Gandhiji* felt very ashamed}.”

Particles: Particles such as भी, ना act as discourse connectives. भी is an emphatic inclusive particle used to suggest the inclusion of verbs, entities, adverbs, and adjectives. Instances of such particles which indicate the inclusion of verbs are taken as discourse connectives (Ex. 3) while others are not.

3) लोग इसे दोनों देशों के बीच बढ़ते रिशते के परिणाम के रूप में देख रहे हैं]।{कशमीरी लोग इससे एक राजनीतिक सबक} **भी** {ले रहे हैं}।

“[People see this as a consequence of the improving relation between the two countries]. {The *Kashmiris* are} **also** {learning an political lesson from this}.”

2.2 Arguments of Discourse Relations

In the PDTB, the assignment of the Arg1 and Arg2 labels to a discourse relation’s arguments is syntactically driven, in that the Arg2 label is assigned to the argument with which the connective was syntactically associated, while the Arg1 label is assigned to the ‘other’ argument. In HDRB, however, the Arg1/Arg2 label assignment is semantically driven, in that it is based on the “sense” of the relation to which the arguments belong. Thus, each sense definition for a relation specifies the *sense-specific semantic role* of each of its arguments, and stipulates one of the two roles to be Arg1, and the other, Arg2. For example, the ‘cause’ sense definition, which involves a causal relation between two eventualities, specifies that one of its arguments is the cause, while the other is the effect, and further stipulates that the cause will be assigned the label Arg2, while the effect will be assigned the label Arg1. Apart from giving meaning to the argument labels, our semantics-based convention has the added advantage simplifying the sense classification scheme. This is discussed further in Section 3.

2.3 Implicit Discourse Relations

The HDRB annotation of implicit discourse relations largely follows the PDTB scheme. The only difference is that while implicit relations in PDTB are annotated only between paragraph-internal adjacent sentences, we also annotate such relations across paragraph boundaries.

3 Senses of Discourse Relations

Broadly, we follow the PDTB sense classification in that we take it to be a hierarchical classification, with the four top level sense *classes* of “Temporal”, “Contingency”, “Comparison”, and “Expansion”. Further refinements to the top class level are provided at the second *type* level and the third *subtype* level. Here, we describe our points of departure from the PDTB classification. The changes are partly motivated by general considerations for capturing additional senses, and partly by language-specific considerations. Figure 1 reflects the modifications we have made to the sense scheme. These are described below.

Eliminating argument-specific labels: In the PDTB sense hierarchy, the tags at the type level are meant to express further refinements of the relations’ semantics, while the tags at the subtype level are meant to reflect different orderings of the arguments (see Section 2.2). In HDRB, we eliminate these argument-ordering labels from the subtype level, since these labels don’t directly pertain to the meaning of discourse relations.

[2] Subordinators that denote the manner of an action are not discourse connectives, but since such disambiguation is a difficult task, we have decided to annotate subordinators in a later phase of the project.

All levels in the sense hierarchy thus have the purpose of specifying the semantics of the relation to different degrees of granularity. The relative ordering of the arguments is instead specified in the definition of the type-level senses, and is inherited by the more refined senses at the subtype level.

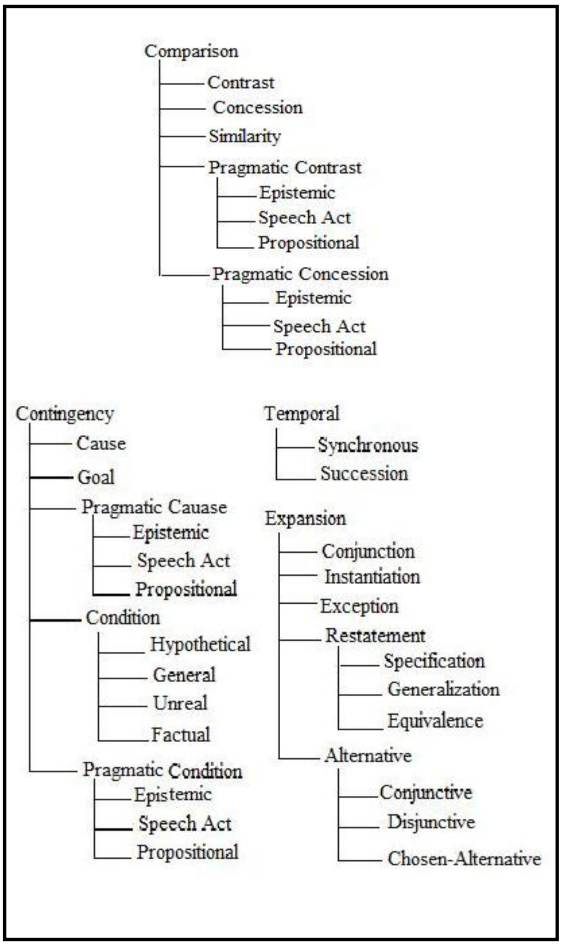

Figure 1: HDRB (Modified) Sense Classification

Uniform treatment of pragmatic relations: As in PDTB, discourse relations in HDRB are pragmatic when their relations have to be inferred from the propositional content of the arguments. However, we replace the PDTB pragmatic senses with a uniform three-way classification. Each pragmatic sense at the type level is further distinguished into three subtypes: "epistemic" (Sweetser 1990), "speech-act" (Sweetser 1990), and "propositional". The propositional subtype involves the inference of a complete proposition. The relation is then taken to hold between this inferred proposition and the propositional content of one of the arguments.

The "Goal" sense: Under the "Contingency" class, we have added a new type "Goal", which applies to relations where the situation described in one of the arguments is the goal of the situation described in the other argument (which enables the achievement of the goal).

4 Initial Annotation Experiments

Based on the guidelines as described in this paper, we annotated both explicit and implicit relations in 35 texts (averaging approx. 250 words/text) from the HDRB corpus. A total of 602 relation tokens were annotated. Here we present some useful distributions we were able to derive from our initial annotation, and discuss them in light of cross-linguistic comparisons of discourse relations.

Types and Tokens of Discourse Relations: Table 1 shows the overall distribution of the different relation types, i.e., Explicit, AltLex, Implicit, EntRel, and NoRel. The second column reports the number of unique expressions used to realize the relation – Explicit, Implicit and AltLex – while the third column reports the total number of tokens and relative frequencies.

Relations	Types	Tokens (%)
Explicit	49	189 (31.4%)
Implicit	35	185 (30.7%)
AltLex	25	37 (6.14%)
EntRel	NA	140 (23.25%)
NoRel	NA	51 (8.5%)
TOTAL	**109**	**602**

Table 1: Distribution of Discourse Relations

These distributions show some interesting similarities and differences with the PDTB distributions (cf. Prasad et al., 2008). First, given that Hindi has a much richer morphological paradigm than English; one would have expected that it would have fewer explicit connectives. That is, one might expect Hindi to realize discourse relations morphologically more often than not, just as it realizes other syntactic relations. However, even in the small data set of 602 tokens that we have annotated so far, we have found 49 unique explicit connectives, which is roughly half the number reported for the 1 million words annotated in English texts in PDTB. It is expected that we will find more unique types as we annotate additional data. The relation type distribution thus seems to suggest that the availability of richer morphology in a language doesn't affect connective usage. Second, the percentage of AltLex relations is higher in HDRB – 6.14% compared to 1.5% in PDTB, suggesting that Hindi makes greater usage of non-connective cohesive links with the prior discourse. Further studies are needed to characterize the forms and functions of AltLex expressions in both English and Hindi.

Senses of Discourse Relations: We also examined the distributions for each sense class in HDRB and computed the relative frequency of the relations realized explicitly and implicitly. Cross-linguistically, one would expect languages to be similar in whether or not a relation with a particular sense is realized explicitly or implicitly, since this choice lies in the domain of semantics and inference, rather than syntax. Thus, we were interested in comparing the sense distributions in HDRB and PDTB. Table 2 shows these distributions for the top class level senses. (Here we counted the AltLex relations together with explicit connectives.)

Sense Class	Explicit (%)	Implicit (%)
Contingency	57 (58.2%)	41 (41.8%)
Comparison	68 (76.5%)	21 (23.5%)
Temporal	43 (65.2%)	23 (34.8%)
Expansion	64(40%)	94(60%)

Table 2: Distribution of Class Level Senses

The table shows that sense distributions in HDRB are indeed similar to those reported in the PDTB (cf. Prasad et al., 2008). That is, the chances of "Expansion" and "Contingency" relations being explicit are lower compared to "Comparison" and "Temporal" relations.

5 Summary and Future Work

This paper has reported on the Hindi Discourse Relation Bank (HDRB) project, in which discourse relations, their arguments, and their senses are being annotated. A major goal of our work was to investigate how well the Penn Discourse Treebank (PDTB) and its guidelines could be adapted for discourse annotation of Hindi texts. To a large extent, we have successfully adapted the PDTB scheme. Proposed changes have to do with identification of some new syntactic categories for explicit connectives, and some general and language-driven modifications to the sense classification. From our initial annotations, we found that (a) there doesn't seem to be an inverse correlation between the usage frequency of explicit connectives and the morphological richness of a language, although there does seem to be an increased use of cohesive devices in such a language; and (b) sense distributions confirm the lack of expectation of cross-linguistic "semantic" differences. Our future goal is to complete the discourse annotation of a 200K word corpus, which will account for half of the 400K word corpus being also annotated for syntactic dependencies. We also plan to extend the annotation scheme to include attributions.

Acknowledgements

This work was partially supported by NSF grants EIA-02-24417, EIA-05-63063, and IIS-07-05671.

References

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for Indian languages. *Proc. of IJCNLP-2008.*

Lucie Mladová, Šárka Zikánová and Eva Hajičová. 2008. From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. *Proc. of LREC-2008.*

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. *Proc. of LREC-2008.*

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

Eve Sweetser.1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* . Cambridge University Press.

Nianwen Xue. 2005. Annotating Discourse Connectives in the Chinese Treebank. *Proc. 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. *Proc. of IJCNLP-2008.*