Chapter 9

Conclusion and Future Work

My thesis is on developing a pilot dependency Treebank for Hindi (HDT) [21]. Treebanks are required for Indian languages. Indian Languages are morphologically richer with relatively flexible word order [83, 115, 36]. Phrase Structure [59] based grammatical models don’t seem to suit these languages very well, whereas Paninian Grammatical Formalism which is a dependency based grammar, suits well to Indian languages. Paninian grammar framework [36] offers a simple syntactico-semantic model. We stated the motivation behind following the Paninian framework in the Indian Language scenario (Refer Chapter 3). We discussed the basic scheme along with some complex Hindi constructions in Chapter 4 and 5 respectively.

I have also showed how a resource such as verb frames lexicon is highly useful as a support resource both for *treebank* annotation and parsing. During the course of developing a scheme for *treebank* annotation I also developed *verb frames* for Hindi (Chapter 6). Around 687 verb frames were developed which show the basic argument structure of these verbs. I have also classified the Hindi verbs based on their argument structure. I have expressed the argument structure of the verbs in terms of karaka relations given in Paninian grammar framework [36]. I have classified the verbs into 49 classes. In the *verb frames*, I have also given the theta role information corresponding to the karaka relations. The mapping of theta roles with karaka relations can be done in the future. I have also worked on the mapping between Propbank arguments and karaka relations and also discussed about Hindi PropBank. In this work, I have taken the Propbank annotation and dependency annotation [21, 36] done on the Hindi data and compared both the annotations to get a mapping between them.

In Section 7.1, I have presented a detailed study of Causative verbs in Hindi wherein I fleshed out the linguistic devices that work for causativization. I gave the classification of causative verbs and the linguistic model that was followed for their classification. The insights on causative verbs have been incorporated in the Hindi dependency treebank [42, 188]. I have also worked on the *verb frames* of the causative verbs. We also plan to use the *verb frames* in a Hindi dependency parser [25] to improve its performance [20]. In this study, I have formed a theoretical background for the study of Hindi causative verbs and say that the implementation of the causative verbs is computationally possible. In the implementation of the causative verbs, the causative feature of a verb can be reflected in the morph analysis. The implementation of this work is again a part of future research work.

In Section 7.2, we discussed the conjunct verbs, a type of complex predicate which are highly productive in Indian languages. In our study of the conjunct verbs, we analyzed some of the diagnostics for manual identification of conjunct verbs and their relevance in automatic identification of conjunct verbs in Hindi. We successfully showed the importance of these diagnostics in statistical techniques by observing the significant increase in overall accuracy of identifying conjunct verbs and their positive effect on parsing accuracy. In future we will try to automate behavioural diagnostics (like Coordination test (D1) and Relativization (D3)) on a large corpus. Although some diagnostics like Constituent Response Test (Wh-Questions) cannot be automated, they can give some theoretical grounding to...
conjunct verb identification and can complement the statistical tool. We tried to include some contextual information through features like f7, but it didn’t help much. Since, additional context proves helpful in many tasks, we will have to explore this feature further. As mentioned, the parsing accuracy shows an improvement when incorporating the features given by our tool. The usefulness of this diagnostic tool can also be tried and tested on other NLP application tasks such as Machine Translation.

We also showed the results of some experiments conducted on the HyDT data which showed that there is a strong correlation between vibhakti-karaka relations. We have reported experiments done by others using the HDT Treebank data (Hyderabad dependency Treebank for Hindi) [21] and also error analysis.

If we look at the overall work, we can conclude by saying that this work of developing a pilot dependency Treebank for Hindi (HDT) [21] has been useful in following ways:

1) It has laid foundation for developing Hindi/Urdu Treebank
2) It has been used in Parsing
3) It has again been a base for developing other Indian Treebanks.
4) Hindi Dependency Treebank has been converted to Universal Dependency (UD) Treebank and this task of converting from Panini to UD could be easily achieved.

In the section given below, I briefly discuss about the work completed. The scope of this thesis work is limited only to the ‘Work Completed’ which is mentioned below under the section 9.1. But this work opens an option for broader research which can be done in the future and does not come under the scope of this thesis. I have listed the future research which can be done, under the ‘Future Work’ section given below in section 9.2

9.1 Work Completed

The following work has been done:

(a) Developed pilot Dependency Treebank for Hindi (HyDT) [21] containing around 2230 sentences.

(b) Developed dependency annotation scheme for Treebank annotation for Hindi.

(c) Validated data of multi-layered and multi-representational treebank for Hindi and Urdu [42, 188].
(d) Analyzed some complex Hindi constructions. There were many more constructions that were analyzed out of which the major ones are covered in this thesis.

(e) During the process of developing pilot Dependency Treebank for Hindi (HyDT) [21], I felt the need of creating Verb frames so I created a supplementary resource of verb frames for 687 Hindi verbs, as it will help in annotating the data.

(f) Classified Hindi verbs based on their verb frames (karaka relations).

(g) HDT (1800 sentences out of 2230) was released for a shared task on Indian Language parsing in ICON-2009 as HyDT-Hindi (Hyderabad Dependency Treebank for Hindi).

(h) Studied Causative verbs in Hindi: (i) Classified causative verbs; (ii) Developed verb frames for Hindi causative verbs; (iii) Created morphological rules for the formation of Hindi causative verbs.

(i) Studied conjunct verbs: Analyzed some of the diagnostics for manual identification of conjunct verbs; Worked on automatic identification of conjunct verbs.

(j) I have also reported the use of HyDT data (Dependency Treebank for Hindi) in various experiments [21] and also reported the error analysis.

(k) I have also worked on the mapping between Propbank arguments and karaka relations.

9.2 Future Work

The following work can be done as future research work (not part of this thesis):

(a) In the supplementary resource of verb frames that I created for 687 Hindi verbs, I have given the theta role information corresponding to the karaka relations in the verb frames. The mapping of theta roles with karaka relations can be done as future work.

(b) I worked on the verb frames of the Hindi causative verbs. In future work, we plan to use all the verb frames in a Hindi dependency parser [25] to improve its performance [20]. The automatic linguistic analysis of the causative verbs is computationally possible. In the implementation of the causative verbs, the causative feature of a verb can be reflected in the morph analysis which again is planned for future work.
(c) In the study of *conjunct* verbs, in future we will try to automate behavioural diagnostics (like D1 and D3) on the availability of large corpus. Although some diagnostics like Constituent Response Test (Wh-Questions) cannot be automated, they can give some theoretical grounding to conjunct verb identification and can complement the statistical tool. We tried to include some contextual information through features like f7, but it didn’t help much. We will have to explore this feature further. The usefulness of this diagnostic tool can also be tried and tested on other NLP application tasks such as Machine Translation.

(d) The HDT Treebank (Hindi Dependency Treebank) can be used in future for creating other Indian languages.

(e) The HDT Treebank (Hindi Dependency Treebank) needs to be explored more to see whether Treebanks developed using this scheme provide finer semantic knowledge.

(f) We also want to explore if the Parsers developed using this scheme prove more effective for complex NLP tasks. Eg: Machine Translation for divergent languages, etc.

(g) The work on Hindi verb frames also need to be explored further theoretically in order to see if this kind of classification can work for other Indian languages.