Building the Uppsala Hindi Corpus

Anju Saxena, Pranava Swaroop Madhyasta and Joakim Nivre
Uppsala University, Department of Linguistics and Philology
{anju.saxena,joakim.nivre}@lingfil.uu.se
pmadhyastha@acm.org

1 Introduction

The aim of this presentation is to describe our on-going work on the building of a written Hindi treebank, referred to here as the Uppsala Hindi Corpus. The Uppsala Hindi Corpus is a part of our project on the construction of a Hindi-Swedish-English parallel treebank, financed by the Swedish Research Council and the Faculty of Languages, Uppsala University.

India is making progress at a rapid pace in the global economy, increasing the need for speedy translation of material as well as a better understanding of its official language (Hindi). Hindi also provides a good testing ground for language technology tools. There are, at present, two major Hindi corpora generally available: the EMILLE corpus (Hardie et al., 2006)[1] and the Hindi corpus at IIT Bombay.[2] The IIT Bombay Hindi corpus is a monolingual corpus, consisting of excerpts of texts each around 2,000 words, and does not seem to have any linguistic annotation. The EMILLE corpus has both a large monolingual and a smaller parallel corpus part (Hindi-English), but this corpus, too, is not linguistically annotated. Given that our goal is to create a trilingual parallel corpus that is POS-tagged and dependency parsed as well as aligned at the word and sentence levels, and that can be used in linguistic research and teaching, these existing Hindi corpora are less useful for our purposes. In the remaining presentation we will focus on our work on the building of the Uppsala Hindi Corpus.

The Hindi materials that, at present, are included in our corpus are:

1. Bible texts: the four Gospels (87,332 words in the Hindi version; Matthew 24,490; Mark 15,400; Luke 26,637; John 20,805)
2. Texts from the parallel corpus section of the EMILLE project (12 300 words)
3. The UN Declaration of Human Rights (1 917 words in the Hindi version)
4. A Hindi novel

Examples in this presentation will be from the Declaration of Human Rights text.

2 Analysis

In this section we will briefly describe the various stages of our work with the Uppsala Hindi Corpus.

2.1 Preprocessing

Some texts were typed in manually for lack of usable electronic versions. All texts have undergone semi-automatic cleaning and converting in order to make them conform to the requirements of the annotation tools used. The texts are now available in XML with Unicode (UTF-8) character representation.

2.2 POS Tagging, Morphological Analysis, and Chunking

POS tagging, morphological analysis and chunking were done using tools developed at IIIT Hyderabad,[3] a morphological analyzer and a version of a shallow parser, modified at Uppsala.

2.3 Addition of NULL Nodes

Hindi frequently uses ellipsis, where not every constituent needs to appear obligatorily in each sentence. Both nouns and verbs can be omitted. For this reason, we have added NULL nodes (or

[1] http://www.emille.lancs.ac.uk/

[2] http://www.cfilt.iitb.ac.in/

[3] http://ltrc.iiit.ac.in/

"empty words"). The following algorithm is used for inserting NULLs:

1. NULL-VG: Insert a NULL verb group (VG) when a sentence does not end with a VG.
2. NULL-CC: Insert a NULL coordinating conjunction (CC) when there is a main finite verb (VFM) in non-sentence final position and this VFM is not followed by a CC.
3. NULL-NN: Insert a NULL noun if there is a chunk with a quantifier (QF) but the chunk does not contain a noun (NN). (Other instances of NULL NN's are beyond the scope of our present analysis.)

We are experimenting with a rule-based approach to the evaluation of NULL node insertion, although this work has been just started. After analyzing the first data set, we hope to find a smoother way of evaluating NULL node insertion.

2.4 Conversion into CoNLL-H Format

The data is then converted from the Shakti Standard Format (SSF) used by the IIIT Hyderabad tools, into the CoNLL-H format needed for the processing of the dependency structure evaluation.[4] In this text-based format, each chunk is described on one line, containing the following tab-separated fields:

- ID: Unique identifier
- C-FORM: Chunk word forms (concatenated)
- H-FORM: Head word form
- C-POSTAG: Chunk POS tags (concatenated)
- H-POSTAG: Head POS tag
- FEATS: Morphological features
- HEAD: ID of syntactic head
- DEPREL: Dependency relation to head

2.4 Dependency Parsing

MaltParser (Nivre et al., 2006) has been trained on a syntactically annotated corpus that is being developed at IIIT in Hyderabad. The parser has then been used to parse the Uppsala Hindi Corpus. The output of this process is the addition of the HEAD and the DEPREL fields in the CoNLL-H format. Not surprisingly, morphological features turn out to be very important for disambiguation in Hindi, which means that the parser makes extensive use of information in the FEATS field in the input.

2.5 Evaluation

At present, we are evaluating and manually correcting the linguistic annotation, both that produced by the IIIT Hyderabad tool chain (POS tagging, morphological analysis, and chunking) and that of the Uppsala tools (NULL node insertion and dependency parsing).

3 Discussion

At present the size of the corpus is small. However, keeping in view the scarcity of language technology resources for Hindi needed for the corpus building (for example, even the unavailability of OCR for Hindi the output of which could have UTF-8 font representation), a decision was made to first concentrate our attention on a relatively small corpus where we test and adjust the various resources available, and then apply this knowledge to a larger size corpus. At each stage of our work a lot of time was spent on checking the results generated to make sure that there are not too many "bugs".

References

Buchholz, S. and E. Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL)*, 149–164.

Hardie, A, P. Baker, A. McEnery, B.D. Jayaram. 2006. Corpus-building for South Asian languages. In A. Saxena, and L. Borin (eds.). *Lesser-Known Languages in South Asia: Status and Policies, Case Studies and Applications of Information Technology.* Mouton de Gruyter.

Nivre, J., J. Hall and J. Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 2216–2219.

Nivre, J., J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel and D. Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 915–922.

[4] The CoNLL-H format is a variant of the CoNLL format developed for multilingual evaluation of dependency parsers (Buchholz and Marsi, 2006; Nivre et al., 2007).