An Approach towards Construction and Application of Multilingual Indo-WordNet

Manish Sinha
Department of Computer Science and Engineering
Indian Institute of Technology Bombay,
Mumbai, India
manish@cse.iitb.ac.in

Mahesh Reddy
Department of Computer Science and Engineering
Indian Institute of Technology Bombay,
Mumbai, India
mahesh@cse.iitb.ac.in

Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay,
Mumbai, India
pb@cse.iitb.ac.in

Abstract

In the work reported here, we present three important related issues.

1. We present an effective method of construction of the Marathi WordNet (http://www.cfilt.iitb.ac.in/wordnet/webmwn/) using the Hindi WordNet (http://www.cfilt.iitb.ac.in/wordnet/webhwn/), both of which are being developed at IIT Bombay. Henceforth we will refer to them as MWN and HWN respectively.
2. The Synset identity is the key to connect WordNets.
3. We present an interface to browse linked Hindi and Marathi WordNets (Bilingual WordNet) simultaneously for a given word either in Hindi or in Marathi.

As an application, we present Word Sense Disambiguation (WSD) of nouns in Hindi. The system has been evaluated on the Corpora provided by Central Institute of Indian Languages (http://www.ciil.org/) and the results are encouraging.

1 Introduction

English WordNet (Fellbaum, 1998) has been used in numerous natural language processing tasks like word sense disambiguation (Agirre 1996; Ramakrishanan 2002; Ramakrishanan 2003), information extraction (Ramakrishanan 2002) and so on with considerable success. Several European language WordNets are connected to form a multilingual lexical knowledge base called EuroWordNet (Vossen 1996). MultiWordNet (Bentivogli 2002) is a project of aligning Princeton's WordNet version 1.6 with Italian WordNet. Hindi (Chakrabarti 2002) and Marathi WordNets are being built at IIT Bombay. Development of Multilingual Web-Scale Language Resources is being carried out as MEANING[1] project. The motivations of such projects are WSD, Cross Lingual Information Processing and large scale knowledge acquisition.

This paper also describes the application of Hindi WordNet to one of the fundamental issues in NLP, *viz.*, Word Sense Disambiguation for Hindi. A supervised learning approach for WSD has been proposed by Yarowsky (Yarowsky 1992). Word associations are recorded and for the unseen text, correct word senses are detected from the learnt associations. Agirre and Rigau (Agirre 1996) use a measure based on the proximity of the text word to a sense in the WordNet (Conceptual Density) to disambiguate the words. The idea that translation presupposes word sense disambiguation is leveraged by Ide (Ide 1999) to disambiguate bilingual corpora. The design of well-known work-bench for sense disambiguation WASP is given by Kilgarriff (1998). Lesk (Lesk 1996) and Lin (Lin 1998) have studied theoretical definitions of similarity and provided word similarity measures that are hypernymy based and gloss based respectively. The notion of soft word sense disambiguation (Ramakrishanan 2003) attempts to rank the senses with a score. The first attempt at Hindi Word Sense Disambiguation was made by Sinha *et. al.* (Sinha 2004).

[1] MEANING (2005) 2nd Workshop organized by MEANING project, Italy http://tcc.itc.it/events/meaning2005/

2 Towards Multilingual Indo WordNet

Relation borrowing in WordNets refers to relation establishment for one WordNet using the relations of another WordNets. The technique is automatic for semantic relations, but semi-automatic for lexical relation. Different cases of relation borrowing from Hindi WordNet (HWN) to Marathi WordNet (MWN) are as follows.

a. *When the sense is present in both Hindi and Marathi:* The relations are established MWN for that sense. This is the commonest case, since Hindi and Marathi belong to the same linguistic family (Indo-Aryan) and exist in almost identical cultural setting.

b. *When the sense is in Hindi but not in Marathi:* The relations will not get established for that sense. For Instance, {दादा [daadaa, grandfather], बाबा [baabaa, grandfather], आजा [aajaa, grandfather], दद्दा [daddaa, grandfather], पितामह [pitaamaha, grandfather], प्रपिता [prapitaa, grandfather]} is a sense in Hindi for paternal grandfather but in Marathi the corresponding sense does not exists.

c. *When the sense is not in Hindi but in Marathi:* The relations for that sense in Marathi have to be established manually. Example {गुढीपाडवा [gudhipaadvaa, newyear], वर्षप्रतिपदा [varshpratipadaa, new year]} is a sense in Marathi which does not have any correspondence in Hindi.

The basic idea of the approach is illustrated in figure 1. The browsable-searchable interface of each of the two WordNets contains the following data structures:

i. A table called *tbl_all_words* which stores for each word the PoS and array of all synset id. For example, for "कर", the table 1 shows the PoS as noun and verb and the synset ids as 491, 3529, 4107, 13314, 13322, 3295, 11958, 11959, 11960, 11961, 11962.

hindi_synset_id	*Word*	*pos*
491	कर	noun
3295	कर	verb
3529	कर	noun
4107	कर	noun
13314	कर	noun
13322	कर	noun
11958	कर	verb
11959	कर	verb
11960	कर	verb
11961	कर	verb
11962	कर	verb

Table 1: HWN.tbl_all_words

marathi_synset_id	*Word*	*pos*
4107	कर	noun
4115	कर	verb

Table 2: MWN.tbl_all_words

ii. A table called *tbl_all_synsets* (table 3 & 4) which stores the sysnet id, the synset and the gloss of the sense.

hindi_synset_id	*Synset*	*gloss*	*category*
491	<not	<not	noun
3295	shown due	shown due	verb
3529	to space	to space	noun
4107	constraint>	constraint>	noun
13341			noun
13322			noun
11958			verb
11959			verb
11960			verb
11961			verb
11962			verb

Table 3: MWN.tbl_all_synsets

marathi_ synset_id	*Synset*	*gloss*	*category*
4107 4115	<not shown due to space constraint >	<not shown due to space constraint >	noun verb

Table 4: MWN.tbl_all_synsets

iii. A table *tbl_<PoS>_<Relation>* for each PoS and Relation combination. For example, *tbl_noun_hypernymy* is the table for the hypernym semantic relation. Continuing the example for "कर", we see that the table 5 stores the various hypernyms of "कर".

synset_id	*hypernymy_id*
491	503
3529	985
4107	3051
13341	12149
13322	1070
11958	2015
11959	3666
11960	7120

Table 5: HWN.tbl_noun_hypernymy

The idea now is to fill the last mentioned tables for MWN *automatically* by following the pointers in the HWN tables. This saves lot of manual effort. Figure 2 describes the algorithm to do this job.

We have also designed a browsable bilingual interface. The input to this browser is a search string in any of the two languages. The browser displays search results for both the languages. The primary usage of this interface is to help users get the semantic information of the search string in both the languages.

```
for each synset identity marathi_synset_id
in Marathi WordNet do
  if (marathi_synset_id == hindi_synset_id)
  do
   for each relation r pointed by
    hindi_synset_id do
    if (relation type of r is semantic) do
     clamp the synset identity linked by
     relation r in to marathi_synset_id
    end if
    else
     clamp the synset identity linked by
     relation r in hindi_synset_id to
     marathi_synset_id AND manually insert
     the corresponding lexical element
    end else
   end for
  end if
end for
```

Figure 1: Relation establishment of Marathi WordNet using Hindi WordNet

3 An Application of HWN: Word Sense Disambiguation

Following Lesk (1996), we give an intersection similarity based WSD approach for Hindi WSD. For the word to be disambiguated, we call the collection of words from the CONTEXT of the word in question the *context Bag* and the related words from the WORDNET as the *Semantic bag*. Figure 2 gives the pictorial view of the approach. Figure 3 gives the algorithm we used.

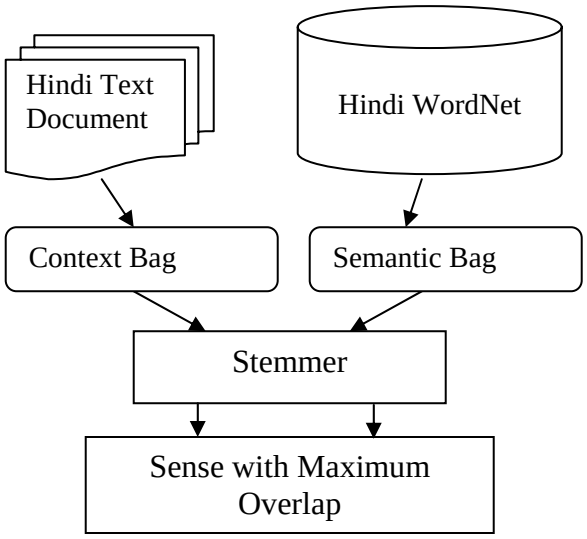

Figure 2: The Basic idea of the WSD approach

1. For a polysemous word ***w*** needing disambiguation, a set of context words in its surrounding *window* is collected. Let this collection be *C*, *the context bag*. The *window* is the current sentence and the *preceding* and the *following* sentences.

2. For each sense *s* of ***w***, do the following. Let *B* be the bag of words obtained from the
   a. Synonyms in the synsets
   b. Glosses of the synsets
   c. Example Sentences of the synsets
   d. Hypernyms (recursively up to the roots)
   e. Glosses of Hypernyms
   f. Example Sentences of Hypernyms
   Hyponyms (recursively up to the leaves)
   g. Glosses of Hyponyms
   h. Example Sentences of Hyponyms
   i. Meronyms (recursively up to the beginner synset)
   j. Glosses of Meronyms
   k. Example sentences of Meronyms

3. Measure the overlap between *C* and *B* using intersection similarity.

4. Output that sense as the winner sense which has the maximum overlap similarity value.

Figure 3: Lesk like algorithm for WSD

We have used Intersection similarity to measure the overlap. The idea of Intersection similarity is to capture the belief that there will be high degree of overlap between the words in the context and the *related words* extracted from the WordNet for a sense, and that sense will be a winner sense.

3.1 Evaluation

We perform Hindi WSD experiment on the corpora provided by Central Institute of Indian Languages (CIIL), Mysore. Currently, the system has been tested on nouns. The domains of the experiment and the accuracy values are mentioned in the table 6. The histogram of figure 4 too shows the WSD accuracy across domains.

Domain	Percentage of Accuracy
Agriculture	73.20
Science and Sociology	64.40
Sociology	60.22
Short-Story	42.22
Mass-Media	50.11
Children Literature	40.00
History	44.44
Science	65.00
Economics	40.55

Table 6: WSD accuracy across domains for Hindi Words

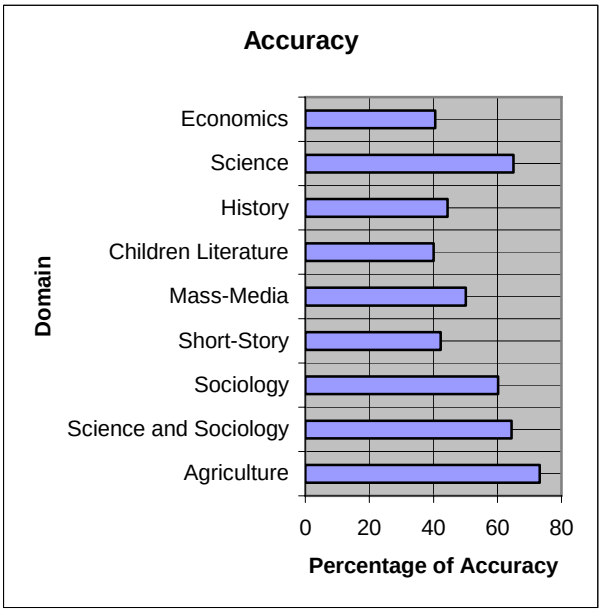

Figure 4: Histogram showing the WSD accuracy across domains for Hindi Words

4 Conclusions and Future Work

In this paper we have described a technique of using Hindi WordNet for establishing Marathi WordNet relations. This is very useful considering the time and effort needed in creating WordNets. To our knowledge, this is the first attempt at linking two Indian language WordNets.

One of the aims of developing the Hindi WordNet was WSD for Hindi. Currently, the system works for Hindi nouns. Work is on for

other parts of speech. Indian Languages are rich in morphology; hence exhaustive stemming is still needed to serve as the front end for the WordNets.

Acknowledgements

This research and development was supported by a grant from the *Ministry of Information and Communication Technology*, Govt. of India. We also acknowledge the WordNet Group (Satish Dethe, Prabhakar Pandey, Sushant Devlekar, Shraddha Kalele, Laxmi Kashyap and Madhura Bapat) at IIT Bombay for their support.

References

Agirre E. and Rigau G. (1996) *Word Sense Disambiguation using Conceptual Density*. COLING, Denmark

Bolksma, L., P. Diez-Orzas, P. Vossen (1996) *User requirements and functional specification of the EuroWordNet project, EuroWordNet*. University of Amsterdam

Chakrabarti D., Narayan D., Pandey P., Bhattacharyya P. (2002) *An Experience in Building the Indo-WordNet- A WordNet for Hindi.* 1st Global Wordnet Conference, Mysore, India.

Fellbaum C. (1998) *WordNet: An electronic lexical database.* MIT press.

Ide. Nancy (1999) *Parallel Translation as Sense Discriminator*. In Proceedings of SIGLEX, University of Maryland, College Park, USA

Kilgarriff A. (1998) *Golden standard Data-sets for Evaluating Word Sense Disambiguation Programs*. In Computer Speech and Language 12(4), Special Issue on Evaluation

Lesk M. E. (1996) *Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a pine cone from an Ice Cream Cone.* SIGDOC.

Lin D. (1998) *An Information-Theoretic Definition of Similarity.* International Conference on Machine Learning, Madison.

Pianta E., Bentivogli M., Girardi C. (2002) *MultiWordNet: developing an aligned multilingual database*. International Wordnet Conference, Mysore, India.

Ramakrishanan G., Bhattacharyya P. (2002) *Word Sense Disambiguation using Semantic Nets based on WordNet*. LREC, Spain.

Ramakrishanan G., Bhattacharyya P. (2002) *Using WordNet based Semantic Sets for Word Sense Disambiguation and Keyword Extraction*. KBCS, India

Ramakrishanan G., Deepa P., Prithviraj B., Bhattacharyya P. and Chakrabarti S. (2003) *Soft Word Sense Disambiguation*. 2nd Global Wornet Conference, Brn, Czeck Republic.

Ramakrishnan G., Prithviraj B., Bhattacharyya P. (2004) *A Gloss Centered Algorithm for Word Sense Disambiguation*. Proceeding of ACL SENSEVAL, Spain.

Resnik P. and D. Yarowsky (2000) *Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation.* Natural Language Engineering, 5(2).

Sinha M., Reddy M., Pande P., Kashyap L., Bhattacharyya P. (2004) *Hindi Word Sense Disambiguation*. International Symposium on MT, NLP and TSS, Delhi, India

Yarowsky D. (1992) *Word Sense Disambiguation using statistical model of Roget's categories trained on large corpora*. COLING, France.