Hindi to English Wordnet Linkage: Challenges and Solutions

Jaya Saraswati Rajita Shukla Ripple P. Goyal Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai – 400076
Maharashtra, India.
{jayas, rajita, ripple, pb}@cse.iitb.ac.in

Abstract

This paper reports the work on linking Hindi wordnet (version 1.2) to the Princeton WordNet (version 2.1), the challenges that were faced while doing so and the solutions to them thereafter. There are a number of concepts common to most of the languages, and linking them with each other can provide an indispensable resource for Natural Language Processing. Hindi wordnet forms the foundation for other Indian language wordnets as those are based on it and are being linked to it. An important strategy of using Direct and Hypernymy linkage to maximize linkages has also been discussed in the paper.

1 Introduction

In a multilingual country like India, machine translation and cross lingual search are highly relevant problems. The wordnets, as crucial linguistic resources, play the most dominant role in the field of text processing applications, such as machine translation, information extraction, information retrieval and natural language understanding systems, and no meaningful research in these areas can be complete without their help. Among the Indian language wordnets, the Hindi wordnet[1] was the first one to come into existence from 2000 onwards. It was inspired by the English WordNet[2] which contains nouns, verbs, adjectives and adverbs organized into synonym sets, each representing one underlying lexical concept (Fellbaum, 1998). Different relations like hypernymy, hyponymy, etc. link the synonym sets to each other. Soon, other Indian language wordnets started getting created. The wordnets for Marathi[3] and Sanskrit[4] followed the Hindi wordnet. All these three efforts are going on at IIT Bombay. Other Indian language wordnets are being linked to the Hindi wordnet, paying particular attention to language specific phenomena. Thus, linking Hindi wordnet to the English WordNet and then linking other Indian language wordnets to Hindi, in turn, will help to increase the linkage of concepts and will create a wide wordnet grid of shared concepts.

2 Roadmap

The roadmap of the paper is as follows: Section 3 presents the background for the linking of the wordnets. Section 4 gives the need for such a linkage while Section 5 and all its sub-sections describe the challenges involved in the task as well as the specific fields where they occurred. Section 6 presents the solutions and Section 7 describes the linking tool and contains its snapshots. Section 8 presents the linkage statistics and section 9 winds up the discussion and gives the conclusion.

3 Scenario in a Multilingual Country like India

India has 22 official languages and hundreds of dialects are spoken across the length and breadth of the nation. The languages of India belong to several linguistic families, the major ones being the Indo-European languages - Indo-Aryan (spoken by 70% of Indians) and the Dravidian languages (spoken by 22% of Indians). Other languages spoken in India belong to the Austro-

[1] http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php
[2] http://wordnet.princeton.edu/
[3] http://www.cfilt.iitb.ac.in/wordnet/webmwn/
[4] http://www.cfilt.iitb.ac.in/wordnet/webswn/

Asiatic, Tibeto-Burman, and a few minor language families and isolates.[5]

Through a number of sponsored projects from the Ministry of Communication and Information Technology under the TDIL program, the wordnets of other languages/language-families started getting created. The first such wordnet was the North East wordnet which was directed at building wordnets for Assamese, Bodo, Manipuri and Nepali (the latter 3 being from Tibeto Burman family). To follow suit were Bangla, Konkani, Gujarati, Punjabi, Kashmiri and Urdu - 6 languages and 7 institutes coming together under the aegis of the Indradhanush project.

4 Need for Linkage

Cross linking words within synsets facilitates the creation of bilingual dictionaries which can be used for several Natural Language Processing tasks such as Machine Translation and Cross Language Information Retrieval. Further, it has been shown that such aligned and linked wordnets can help to perform Word Sense Disambigutaion even in the absence of sense tagged corpora in a target language (Khapra et al, 2009). This is achieved by projecting various parameters essential for WSD from the tagged corpus of a resource rich source language using these cross linkages as a bilingual bridge. Keeping in view these needs, the wordnets of all Indian languages are to be linked with the Hindi wordnet, while the Hindi wordnet itself is being linked to the English WordNet (J. Ramanand et al, 2007). As the Hindi wordnet has been created from the first principles instead of following the expansion approach of building a wordnet, linkage of the Hindi wordnet to the English WordNet becomes an important task (Narayan et al, 2002).

5 Challenges in Linkage

The linkage task has to do a fine balance between maintaining accuracy and providing maximum linkages. While trying to do this for the linkage between the Hindi wordnet and the English WordNet, several challenges were encountered. The problems have occurred because the two languages belong to cultures and social mores which are widely different. The specific areas where such problems were faced are the synsets denoting the following:

- kinship relations
- musical instruments
- kitchen utensils
- tools
- species
- grains

5.1 Kinship Relations

Most kinship terminologies distinguish between sexes (the difference between a brother and a sister) and between generations (the difference between a child and a parent) and also distinguish between relatives by blood and marriage. Different languages (and, by extension, societies) organize these distinctions differently. Kin terms and terminologies can either be descriptive or classificatory (Morgan, 1871). Societies generally use a combination of the two. When a descriptive term is used, it can only represent one type of relationship between two people, while a classificatory term represents one of many different types of relationships. For example, the word *brother* in English-speaking societies indicates a son of the same parent; thus, English-speaking societies use the word *brother* as a descriptive term. But a person's male first cousin could be the mother's brother's son, mother's sister's son, father's brother's son, father's sister's son, and so on; English-speaking societies therefore use the word *cousin* as a classificatory term.

The classificatory terms of English pose a unique problem when they are used for matching descriptive terms of Hindi. English does not have different terms for various relationships such as for *uncles, aunts, cousins, brother-in-laws, sister-in-laws, grandparents and grandchildren* of the different sexes on the maternal and paternal sides, separately. On the other hand, Hindi has, for example, different terms for paternal uncle (father's brother), maternal uncle (mother's brother) as चाचा (*caacaa*) and मामा (*maamaa*) respectively and for father's sister's husband and mother's sister's husband as फूफा (*aahpoohp*) and मौसा (*mausaa*) respectively. All these terms have only one corresponding English term - *uncle* – (*the brother of your father or mother; the husband of your aunt*).

[5] www.wikipedia.org

5.2 Musical Instrument

Music, along with the other components of fine arts, is closely related to the culture of a land. The terms denoting the various musical instruments are unique to a particular language. Finding an exact match to link them in a language which belongs to a very different culture is a great challenge. The hyponymy relation of the Hindi word *वाद्ययंत्र* (*vaadyayayantra*) gives the names of various musical instruments, for many of which an exact corresponding English term is not available. For example, in Hindi there is a percussion instrument, *तबला* (*tabalaa*) – ताल देने का एक वाद्य, जिसमें दो बाजे एक साथ बजते हैं,

- taala dene kaa eka vaadya, jismein do baaje eka saatha bajate hain
- an instrument to provide beats, where two instruments are played together

which has no corresponding English term.

5.3 Kitchen Utensils

This is yet another category of words which are highly culture-specific, being related to food, modes of cooking and eating habits of people. A look at the hyponymy of the term *बर्तन* (*bartan*) - धातु, शीशे, मिट्टी, आदि का वह आधार जिसमें खाने-पीने की चीज़ें रखी जाती हैं

- dhaatu, sheeshe, mitti, aadi kaa vaha aadhaara jismein khaane-piine kii ceezein rakhii jaatii hain
- a base made of metal, glass, clay, etc. in which eatables are kept

will throw up a long list of terms denoting various utensils, for many of which exact English terms are not present. For example, Hindi words like *डोंगा* (*dongaa*) - भोजन रखने का एक तरह का कटोरा,

- bhojana rakhane kaa eka taraha kaa katoraa
- a type of bowl for keeping food

or *कटोरदान (katoradaan*) - वह ढक्कनदार बर्तन जिसमें भोजन आदि रखते हैं

- vaha dhakkanadaar bartana jismein bhojana aadi rakhate hain
- that lidded pot in which food is kept

are very much culture-specific. Also, there is a problem of size as Hindi distinguishes between terms for big and small of the same object, whereas English has a single term for both. For example, the words *कलछा* (*kalachhaa*) - बड़ी डाँड़ी का चम्मच जिससे बटलोई आदि की दाल आदि चलाते या निकालते हैं

- badii daandii kaa cammaca jisase bataloi aadi kii daala aadi calaate yaa nikaalate hain
- a long-handled spoon which is used to stir or serve lentils etc. from a pot

and *कलछी* (*kalachhii*) - एक छोटा कलछा

- eka chhotaa kalachhaa
- a small spoon

denote the same utensil, the former being big and the latter of a small size. The corresponding English term, *ladle (a spoon-shaped vessel with a long handle; used to transfer liquids)*, is not size-specific.

5.4 Tools

Here again, the problem of exact matches and the differentiation of big and small exists. Terms like *कनखोदनी* (*kanakhodanii*) - कान से मैल निकालने का विशेषकर चम्मच की शक्ल का एक छोटा उपकरण

- kaan se maila nikaalane kaa visheshakara cammaca kii shakla kaa eka chhotaa upakara-Na
- a small spoon-shaped instrument for removing ear wax

or *अंकुसी* (*ankusii*) - नारियल के भीतर गरी निकालने वाला एक औजार जिसका सिरा नुकीला होता है

- naaryala ke bhiitara garii nikaalane wallah eka aujaara jisakaa siraa nukiilaa hotaa hai
- a tool having a pointed end used for removing the meat from a coconut

do not have corresponding English terms. Both the words *खुर्पा* (*khurpaa*) - घास आदि छीलने का एक औजार

- ghaasa aadi chhiilane kaa eka aujaara

- a tool to remove grass, etc.

and *खुर्पी* (*khurpii*) - छोटा खुर्पा

- chhotaa khurpaa

- a small spud

denote the English concept of *spud, stump spud - a sharp hand shovel for digging out roots and weeds*.

5.5 Species

The problem of linkage of terms denoting species of both birds and animals is quite unique in its own self. The Hindi wordnet has different terms for denoting the male and female of the species, besides having a synset for denoting the species itself which is not gender-specific. For example, the synset of *शेर* (*śera*) - बिल्ली की जाति का एक बहुत बड़ा और भयंकर, हिंसक पशु

- billii kii jaati kaa eka bahuta badaa aura bhayaMkara, hinsaka pashu

- a huge, fearsome and savage animal belonging to the cat family

denotes the species *tiger*, the synset *शेर* (*śera*) - शेर जाति का नर (śera jaati kaa nara; male of the tiger species) is for the male tiger and *शेरनी* (*śeranii*) - मादा शेर for the tigress. While it is easy to link the synsets denoting the species and the synsets denoting the female of the species in this case, a separate synset for the male of the species is not separately included in the English WordNet. Some concepts do not have synsets for the male and the female of the species at all and linkage remains a challenge. For example, the English WordNet has a single synset of *frog* (any of various tailless stout-bodied amphibians with long hind limbs for leaping; semi-aquatic and terrestrial species) while the Hindi wordnet has three separate synsets for it - *मेंढक* (*me^{n}ḍhaka*) - एक छोटा बरसाती उभयचर प्राणी जो प्रायः वर्षा ऋतु में तालाबों, कुओं आदि में दिखाई देता है

- eka chhotaa ubhayacara praaNii jo praayaH varshaa ritu mein taalaaboM, kuoM, aadi mein dikhaaii detaa hai

- a small amphibian which is often seen in ponds, wells, etc. during the rainy season

which stands for the species *frog*, the synset *मेंढक* (*me^{n}ḍhaka)* - नर मेंढक for the male frog and *मेंढकी* (*me^{n}ḍhakii*) - मादा मेंढक for the female frog.

5.6 Grains

Here, the problem lies in the synsets of the various types of millets which are a part of the Hindi wordnet, such as *ज्वार* (*jwaara*), *बाजरा* (*baajaraa*), *मडुआ* (*maṛuaa*), etc. For these, it is very difficult to find exact English terms. It has been observed that the standard Hindi-English bilingual dictionaries give the general term *millet* for many of these or different English word for the same Hindi word. In such a scenario it is quite difficult to conclusively state which the correct term is. For example, for the term *बाजरा* (*baajaraa*), the Oxford Hindi-English Dictionary[6] gives the word *millet* while some other dictionaries add to the confusion by giving yet other terms for it.

6 Solution

To find a solution to the above problems, it has been decided to use two kinds of linkages – **Direct** and **Hypernymy linkages**. Synsets having exact equivalents in English WordNet are to be linked through direct linkage. For example, *आम* (*aama*), *आम वृक्ष* (*aama vriksha*) is to be linked to the English synset *mango, mango tree*.

The synsets which cannot be linked directly to English concepts are to be linked through hypernymy. This means that in the absence of the equivalent English concept, the nearest term capturing the sense would be assumed as the hypernymy of that concept and would be linked to it. This would be known as hypernymy linkage. This information would be captured by the linkage tool. For example, the Hindi synsets of *चाचा* (*caacaa*) and *मामा* (*maamaa*) would be linked to the English synset of *uncle* through hypernyny linkage. The same would be applicable for musical instruments where, for example, the synset of *drum* would be assumed to be the hypernymy for the Hindi synset of *तबला (tabalaa)*. Similarly, for kitchen utensils such as *डोंगा* (*dongaa*), the assumed hypernymy would be

[6] R.S. McGregor. *The Oxford Hindi-English Dictionary* (ed.). Oxford University Press, New Delhi, India.

tableware (*articles for use at the table (dishes and silverware and glassware*)) while for *कटोरदान* (*katoradaana*) it would be *container* (*any object that can be used to hold things (especially a large metal boxlike object of standardized dimensions that can be loaded from one form of transport to another*)). The same would be done for objects such as *कनखोदनी* (kanakhodanii) and *अंकुसी* (*ankusii*) which would be linked to their hypernymy, in this case, to *tool* (*an implement used in the practice of a vocation).*

For concepts which have separate synsets for the big and small of the same object, it has been decided to link the concept which is more frequently used as direct linkage with the corresponding English synset, while the one which is not as frequently used as hypernymy linkage. The decision about the most frequently used word would depend on the lexicographer's point of view, native speakers' intuition and Google query about the frequency of usage of the word. For example, between *खुर्पी* (*khurpii*) and *खुर्पा* (*khurpaa*) – *खुर्पी* (*khurpii*) would be linked as direct linkage to *spud* as it is more commonly used, while *खुर्पा* (*khurpaa*) would also be linked to *spud*, but as hypernymy linkage. The same would hold true for *कलछा* (*kalachhaa*) and *कलछी* (*kalachhii*), where both would be linked to *ladle*, with the former to be a hypernymy linkage and the latter as a direct linkage to due to the same reason.

To deal with the problem posed by male-female distinction of species, the solution would be to link the Hindi synset denoting species to the synset of English as direct linkage, while linking the male of the species to the same synset as hypernymy linkage. Thus, the synset *शेर* (*śera)* - बिल्ली की जाति का एक बहुत बड़ा और भयंकर, हिंसक पशु which denotes the species, would be linked to the English synset of *tiger* (*large feline of forests in most of Asia having a tawny coat with black stripes; endangered)* - as direct linkage, while the Hindi synset *शेर* (*śera*)- शेर जाति का नर would be again linked to the same English synset as hypernymy linkage. In the case of मेंढक, both the synsets of नर मेंढक and मेंढकी would be linked to *frog* through hypernymy linkage.

For grains too, the linkage would be done to the assumed hypernymy, for example *ज्वार* (*jwaara*), *बाजरा* (*baajaraa*) would be linked to *millets (small seed of any of various annual cereal grasses especially Setaria italic)* - under hypernymy linkage**.**

7 Linking Tool

The Hindi wordnet uses a semi-automated system for linking the Hindi wordnet with the English WordNet. The WNSynsetMatcher tool, used by lexicographers for manually linking the two wordnets, was developed at CFILT, IIT Bombay. A screen shot of the tool is shown in Figure 1.

The tool takes as input a file containing the number of query synsets N, where N stands for total number of synsets that are to be linked and N lines in following format:

- Source synset ID
- POS category
- Number of candidates synsets in target language
- Candidate synset ID and corresponding confidence score for each candidate.

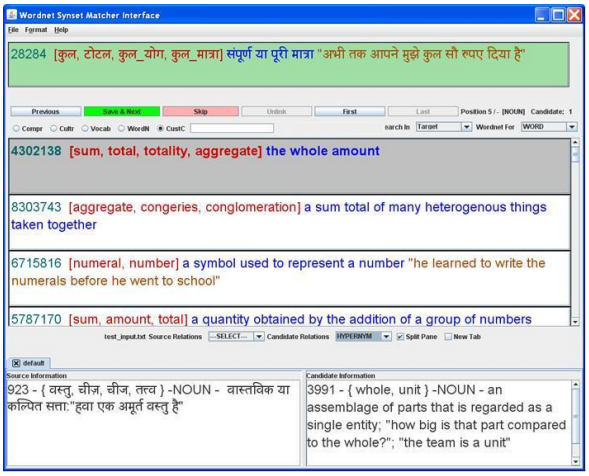

Figure 1. WNSynsetMatcher Tool

The candidate synsets and their confidence scores are to be obtained by applying various heuristics (Karra, 2010).

In the tool, the source synset (synset ID, synonyms, POS category, gloss and example) is displayed in the source synset panel at the top of the tool. Similar information is displayed in the candidate synset panel below it, for each of the *N*

candidate synsets. The candidates are displayed in decreasing order of their confidence score. Facility for searching synsets in both source and target languages with respect to a word or synset ID is also provided in the tool.

The lexical and semantic relations for a given source synset and any of the candidate synsets can be viewed in the bottom Information panel. This panel also supports multiple tabs to facilitate detailed scrutiny of ambiguous candidate synsets.

In case the lexicographer is unable to find a suitable match for the given source synset among the candidate synsets, then he can skip linking this synset by adding standard comments provided as radio buttons. The comments are:

- Comprehension (Compr): difficulty in understanding the concept.
- Culture (Cultr): unavailability of a target synset when the concept described by the source synset is specific to a particular culture (in our case India).
- Vocabulary (Vocab): synset members are out of standard vocabulary.
- Wordnet (WordN): absence of counterpart in English WordNet.
- Customized Comment (CustC): customized comment which can be entered by the lexicographer.

The statistics of the work completed i.e., the numbers of linked and skipped synsets etc., are displayed before the tool is closed.

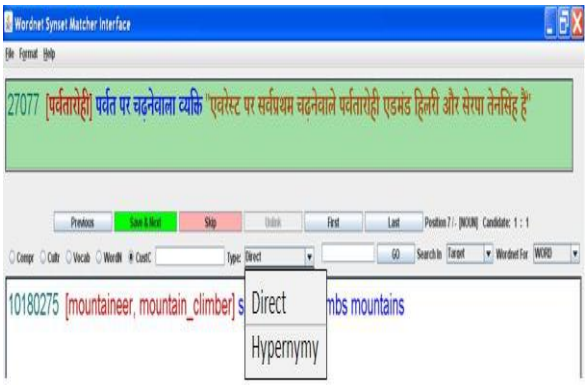

Figure 2. Modified WNSynset Matcher tool

To solve the problems faced by the lexicographers during linkage and maximize the number of synset being linked, the concept of 'type of link' was introduced. To capture this type of relation between the linked synset pair, i.e. Direct or Hypernymy, the tool has been modified. Using a drop down box lexicographers have to explicitly select the nature of the link before saving a pair of synsets. A screen shot of the modified tool can be seen in figure 2.

The work done by the lexicographer is dumped into files on a daily basis. These files are later processed and uploaded on the database. As soon as the data is uploaded, it is immediately reflected on the online version of Hindi wordnet. The online version displays tags (Direct or Hypernymy) to indicate the type of linkage shared as shown in figure 3.

Figure 3. Tags representing the type of link

8 Linkage Statistics

Total Hindi synsets	34419
Number of synsets linked	15062
Number of synsets skipped	13934
Number of synsets left for first consideration	5423
Hypernymy linked	5
Direct linked	15057

Table 1. Current linking statistics (as of 1st November 2010)

9 Conclusion

In this paper we have discussed the linking of the Hindi wordnet to the English wordnet, the challenges therein and the solutions that were found to overcome them. It was observed that the problems occurred due to cultural differences. The main problem areas are the following:

- kinship relations
- musical instrument
- kitchen utensils
- tools
- species

- grains

As a solution the strategy of using Direct and Hypernymy linkages is suggested. The linking tool has been modified accordingly.

It is hoped that by the method described above the task of maximizing linkages can be accomplished with a high degree of accuracy. Lexicographers of the other wordnets of the Indian languages will find it very useful while linking their synsets to the Hindi wordnet. It would be an extremely interesting study to see as to what other cases of the use of the hypernymy linkage come up. Overall, this would bring about an unprecedented, unique conceptual unity among India's many and varied languages.

Acknowledgments
We gratefully acknowledge the support from Department of Technology, Ministry of Communication and Information Technology. We also acknowledge Salil Joshi for modifying the online linkage tool and Mitesh Khapra for his valuable inputs, both from the computational team of CFILT, IIT Bombay.

References

Arun Karthikeyan Karra. 2010. *WordNet Linking.* Master of Technology Dissertation, CSE Department, IIT Bombay.

Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande and P. Bhattacharyya. 2002. *An Experience in Building the Indo WordNet- a WordNet for Hindi*. International Conference on Global WordNet (GWC 02), Mysore, India.

Fellbaum, C. 1998. *Wordnet: An Electronic Lexical Database*. The MIT Press.

J. Ramanand, Akshay Ukey, Brahm Kiran Singh, Pushpak Bhattacharyya. 2007. *Mapping and Structural Analysis of Multi-lingual Wordnets.* IEEE Data Engineering Bulletin, 30(1).

Kamil Bulke. 1997. *An English-Hindi Dictionary* (ed.). S. Chand & Co, New Delhi, India.

Lewis Henry Morgan. 1871. *Systems of consanguinity and affinity of the human family*. Smithsonian Contributions to Knowledge; v. 218, Washington DC.

Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya. 2009. *Projecting Parameters for Multilingual Word Sense Disambiguation.* Empirical Methods in Natural Language Processing (EMNLP09), Singapore.

Dr. S. Awasthi and Dr. (Smt.) I. Awasthi. 2000. *Chambers English-Hindi Dictionary* (ed.). Allied Publisher Limited, New Delhi, India.

www.Shabdkosh.com

www.wikipedia.org

http://pustak.org/bs/home.html

http://www.thefreedictionary.com