Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT*

Debasri Chakrabarti and Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology, Mumbai, India
Email: debasri@cse.iitb.ac.in, pb@cse.iitb.ac.in

Abstract. Verbs form the pivots of sentences. However, they have not received as much attention as nouns did in the ontology and lexical semantics research. The classification of verbs and placing them in a structure according to their selectional preference and other semantic properties seem essential in most text information processing tasks like machine translation, information extraction *etc*. The present paper describes the construction of a verb hierarchy using Beth Levin's verb classes for English, the hypernymy hierarchy of the WordNet and the constructs and the knowledge base of the Universal Networking Language (UNL) which is a recently proposed interlingua. These ideas have been translated into the building of a verb hierarchy for Hindi. The application of this hierarchy to the construction of the Hindi WordNet is discussed. The overall motivation for this work is the task of machine translation between English and Hindi.

1 Introduction

The verb is the binding agent in a sentence. The nouns in a clause link to the main verb of the clause according to the verb's selectional preferences. However, verbs have not received as much attention as they deserve, when it comes to creating lexical networks and ontologies. Ancient Sanskrit treatises on ontology like the *Amarkosha* [1] deal meticulously with nouns, but not with verbs. The present day ontologies and lexical knowledge bases like *CYC [2], IEEE SUMO [3], WordNet [4,5], EuroWordNet [6], Hindi WordNet [7]*, *Framenet [8] etc.* build deep and elaborate hierarchies for nouns, but the verb hierarchies are either not present or if present are shallow. The *Verbnet* project [9] is concerned exclusively with verbs and builds a very useful structure, but does not concern itself with building a *hierarchical structure.*

The classification of verbs and placing them in a structure according to their selectional preference and other semantic properties seem essential in most text information processing tasks [9,10] like machine translation, information extraction *etc*. Additionally, *property inheritance* (*e.g. walk* inherits the properties of *move*) facilitates lexical knowledge building, for example, in a rule based natural language analysis system [11].

* Editor's note: This version of the paper may have not been typeset correctly due to the author's own formatting and non-availability of all fonts needed for retypesetting. The author's version is to be found on the accompanying CD.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 83–90.

The present paper describes the creation of a hierarchical verb knowledge base for an interlingua based machine translation system based on *Universal networking Language (UNL)* [12] and its integration to the Hindi WordNet. Use is made of (i) English verb classes and their alternation [10], (ii) the hypernymy hierarchy of WordNet [4,5] and the specifications and the knowledge base of the UNL system [12].

The organization of the paper is as follows. Section 2 deals with Levin's classification of English verbs. Section 3 is a brief introduction to the UNL system and the verb knowledge base therein. The creation of the verb hierarchy is explained in Section 4 with focus on the Hindi verbs. Section 5is on verbs and the Hindi WordNet. Section 6 concludes the paper and gives *future directions.*

2 Levin's Class of English Verbs

The key assumption underlying Levin's work is that the *syntactic behavior of a verb is semantically determined* [10]. Levin investigated and exploited this hypothesis for a large set of English verbs (about 3200). The syntactic behavior of different verbs was described through one or more alternations. *Alternation describes a change in the realization of the argument structure of a verb*, *e.g. middle alternation, passive alternation, transitive alternation etc*. Each verb is associated with the set of alternations it undergoes. A preliminary investigation showed that there is a considerable correlation between some facets of the semantics of verbs and their syntactic behavior so as to allow formation of classes. About 200 verb semantic classes are defined in Levin's system. In each class, there are verbs that share a number of alternations. Some example of these classes are the classes of the *verbs of putting*, which include *put verbs*, *funnel verbs*, *verbs of putting in a specified direction, pour verbs, coil verbs*, *etc*.

3 The Universal Networking Language (UNL)

The Universal Networking Language (UNL) [12] is an electronic language for computers to express and exchange information. UNL system consists of *Universal words (UW)* (explained below), *relations, attributes*, and the *UNL knowledge base (KB)*. The UWs constitute the vocabulary of the UNL, relations and attributes constitute the syntax and the UNL KB constitutes the semantics. The KB defines possible relationships between UWs.

UNL represents information sentence-by-sentence as a hyper-graph with concepts as nodes and relations as arcs. The representation of the sentence is a hyper-graph because a node in the structure can itself be a graph, in which case the node is called a *compound word (CW)*. Figure 1 represents the sentence *John eats rice with a spoon*.

In this figure, the arcs labeled with *agt* (agent), *obj* (object) and *ins* (instrument) are the relation labels. The nodes *eat(icl>do), John(iof >person), rice(icl>food)* and *spoon(icl>artifact)* are the *Universal Words* (*UW*). These are language words with *restrictions* in parentheses. *icl* stands for *inclusion* and *iof* stands for *instance of*. UWs can be annotated with attributes like *number*, *tense etc.* which provide further information about how the concept is being used in the specific sentence. Any of the three restriction labels- *icl, iof* and *equ-* can be attached to an UW for restricting its sense.

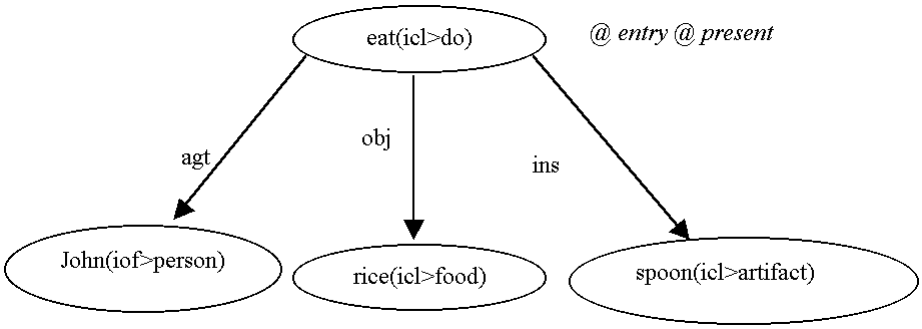

Fig. 1. UNL graph of *John eats rice with a spoon*

3.1 Verbal Concepts in UNL

The verbal concepts in the UNL system are organized in three categories. These are:

(icl>do) for defining the concept of an event which is caused by something or someone.
e.g., change(icl>do): as in *She changed the dress.*

(icl>occur) for defining the concept of an event that happens of its own accord.
e.g., change(icl>occur): as in *The weather will change.*

(icl>be) for defining the concept of a *state verb*.
e.g., remember(icl>be): as in *Do you remember me?*

The first two categories correspond to the *action and* the *event verbs* respectively of the *nonstative class* and the third corresponds to *stative* [13]. A part of the hierarchy for the top concept *do* is shown in Figure 2.

```
do(agt>thing{,^gol>thing,icl>do,^obj>thing,^ptn>thing,^src>thing})
     → do(agt>volitional thing{,icl>do(agt>thing)})
         → do(agt>living thing{,icl>do(agt>volitional thing)})
            → do(agt>human{>living thing,icl>do(agt>living thing)})
     → do(agt>thing,gol>thing{,icl>do,^obj>thing,^ptn>thing,^src>thing})
```

Fig. 2. Partial hierarchical structure for *do*

The semantic hierarchy of the *do* tree is shown in Figure 3.

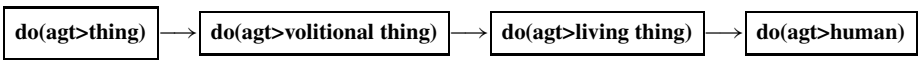

Fig. 3. Semantic hierarchy for *do*

```
"move" 'We should move ahead in this matter.' (to follow a procedure
       or take a course)
   (icl>act(agt>person))"
   [VINTRANS,VOA-ACT]
   → "move" 'How fast does your new car move?'
      (to change location)
      (icl>motion(>act(agt>thing))
      [VINTRANS,VOA-MOTN,VOA-ACT]
      → "move" 'Due to rain the cows were moving fast.'
      (to change the place or position of your body or a part of your body)
      (icl>motion{>act}(agt>volitional thing))
      [VINTRANS,VOA-MOTN,VOA-ACT]
   → "move" 'She cannot move her fingers.'
       (to cause to change the place or position of your body
       or a part of your body)
       (icl>motion(>act}agt>thing,obj>thing))
       [VTRANS,VOA-MOTN,VOA-ACT]

"move" She's made up her mind and nothing will move her.
(to change one's attitude or make sb change their attitude)
(icl>affect{>change}(agt>thing,obj>thing))
[VTRANS,VOO-CHNG]
```

Fig. 4. A part of *move* hierarchy

The specified relations for the *do* category are *agent*, *object*, *goal*, *partner* and *source*. It is stated that *agent* is the compulsory relation for this category. The *do* verb appearing in the hierarchy with only *agt* relation is the top node. In Figure 2, the symbol "^" specifies the *not* relation. It states that the top node of *do* does not take *gol (goal)*, *obj (object)*, *ptn (partner)* and *src (source)* relations. The second node in the figure shows that *do* appearing with *agt* and *gol* relation is the child of the top node. This hierarchy is set up using the argument structure of the verb. In the hierarchy the symbol '→' stands for the *parent-child* relationship.

4 Creation of the Verb Hierarchy

Levin's verb classes form the starting point. All the classes and the sub-classes are then categorized according to the UNL format (*vide* the previous section). Generally, to select the *top node*, the WordNet hypernymy hierarchy is used. However, when the WordNet hierarchy is not deep enough, dictionaries are used to arrive at the top node based on the perceived meaning hierarchy. Figure 5 shows a part of the hierarchy for the verb *put* (Similar partial tree for *move* appears in Figure 4). Everywhere, we first give the name of the verb, followed by an example sentence, the WordNet gloss, the UNL KB representation, the syntax frame and finally the grammatical and semantic categories (VTRANS, VOA-ACT etc.).

This example shows two types of sentence frames for the *put* class: one with the locative preposition (*in, around, into* etc.) and the other with the place adverb frame (*here/ there*). *hang* is the child node of *put*.

```
"put"
'Put your clothes in the cupboard'.
(to put something into a certain place)
(icl>move(agt>person,obj>concrete thing,gol>place)
(loc_prep{in/on/into/under/over)
[VTRANS, VOA-ACT]
   → "hang"
      'He hanged the wallpaper on the wall'.
      (to suspend or fasten something so that it is held up
       from above and not supported from below)
      (icl>put{>move}(agt>person,obj>concrete thing,gol>place)
      (loc_prep{from/on)
      [VTRANS, VOA-ACT]
"put"
'Put your things here'.
(to put something into a certain place)
icl>move(agt>person,obj>concrete thing,gol>place)
adv_plc{here/there)
[VTRANS, VOA-ACT]
```

Fig. 5. Hierarchy of the *put class*

4.1 Verb Hierarchy in Hindi

We elucidate the ideas with the example hierarchy for the Hindi verb रखना **rəkʰna** (rakhanaa, meaning *put*) shown in Figure 6. In this figure, the name of the verb in Hindi is first mentioned, followed by the IPA transcription and the English transliteration. Then the corresponding English verb is given followed by the gloss from the English WordNet. After this comes the UNL representation with the example Hindi sentence (in IPA and English transliteration) and the sentence frame.

It is evident that there is a difference in the syntax frame with respect to English. For example, for the adverbial-place frame in English, the Hindi frame contains a locative postposition. This is due to the fact that case markers are obligatory features in the syntax of Hindi which is an inflectional language.

There are two different syntax frames specified for the *put class* in English [10], *viz., adv plc* and *loc prep*. Hindi has an extra frame for the same class. Thus, the syntax frames for the रखना (*put*) *class* are:

a. adv man
b. adv plc adv man
c. loc postp adv man.

This leads to the discussion on the difference in the representations for *troponyms* in the two languages. *In English, the troponyms of a verb are usually different lexical terms. In Hindi, generally the verb itself with different syntax frames represents the troponyms.* It can thus be inferred that troponyms are lexically specified in English and syntactically in Hindi. The example of *arrange* in figure 7 makes this point clear.

A summary of the syntax frames for the verb *arrange* in the two languages is shown in Table 1.

रखना **rəkʰna rakhanaa**

put' 'Put your things here.' (to put something into a certain place)
(icl act(agt person,obj concrete thing,gol place)
अपना समान यहाँ पर रखो (əpna səman yəhã pər rəkʰo) apanaa samaan yahaa par rakho
(adv plc (यहाँ वहाँ, 'yəhã / vəhã') loc postp (पर, 'pər')

→ रखना, सजाना **rəkʰna, səjana rakhanaa, sa aanaa**

arrange' 'He arranged the books here.' (to put something in a particular order to put into a proper or systematic manner)
(icl put act (agt person,obj thing)
उसने किताबों को यहाँ पर सजाकर रखा। usne kitabõ ko yəhã pər səjakər rəkʰa. usne kitabo ko yahaa par sajaakar rakhaa.
(adv man (सजाकर, 'səjakər' क्रम से, 'krəm se') (adv plc (यहाँ वहाँ 'yəhã / vəhã') loc postp(पर, 'pər')

→ ढेर लगाना, इकट्ठा करना ɖʰer ləgana, ikəʈʈʰa kərna **hera lagaanaa, ikaTThaa karanaa**

heap' 'He heaped woods here.' (to arrange in stacks)
(icl arrange put (agt person,obj functional thing,gol functional thing)
उसने यहाँ पर लकड़ियां इकट्ठा की। usne yəhã pər ləkɖiyã ikəʈʈʰa kı. usne yahaa par lakdiya ikatthaa kii.
(adv plc (यहाँ वहाँ, 'yəhã / vəhã loc postp (पर, 'pər')

Figure 6 Hierarchy for *रखना put*

रखना, सजाना **rəkʰna, səjana rakhanaa, sa aanaa** 'arrange'

a. Sentence: उसने किताबों को सजाकर रखा। usne kitabõ ko səjakər rəkʰa. usne kitabo ko sajaakar rakhaa.
उसने किताबों को क्रम से सजाया। usne kitabõ ko krəm se səjayi. usne kitabo ko kram se sajaayaa.
'He arranged the books.'
Frame: adv man (सजाकर, 'səjakər' क्रम से , 'krəm se')

b. Sentence: उसने यहाँ पर किताबें क्रम से सजायीं सजाकर रखीं।
usne yəhã pər kitabẽ krəm se səjayĩ səjakər rəkʰĩ. usne yahaa par kitabe kram se sajaayii sajaakar rakhii.
'He arranged the books here.'
Frame: adv plc (यहाँ वहाँ, 'yəhã / vəhã') loc postp (पर, 'pər') adv man (सजाकर, 'səjakər' क्रम से, 'krəm se')

c. Sentence: उसने मेज के ऊपर किताबें क्रम से सजायीं सजाकर रखीं।
usne mej ke upər kitabẽ krəm se səjayĩ səjakər rəkʰĩ. usne mej ke uupar kitabe kram se sajaayii sajaakar rakhi.
'He arranged the books on the table.'
Frame: loc postp (के ऊपर, 'ke upər' के नीचे, 'ke nice') adv man (सजाकर, 'səjakər' क्रम से, 'krəm se')

Figure Sentence frames for *arrange*

Table 1 Sentence frames for *arrange*

English	Hindi
1. adv plc (here there)	1. adv man (सजाकर, 'səjakər' क्रम से, 'krəm se' *etc.*)
2. loc prep (in, inside, on *etc.*)	2. adv plc (यहाँ वहाँ, 'yəhã / vəhã') loc postp (पर, 'pər') adv man (सजाकर, 'səjakər' क्रम से, 'krəm se' *etc.*)
	3. loc postp (के ऊपर, 'ke upər' के नीचे, 'ke nice' *etc.*) adv man (सजाकर, 'səjakər' क्रम से, 'krəm se' *etc.*)

5 Verbs and the Hindi WordNet

The differences between Hindi and English verbs give rise to *language divergences* in machine translating one language to the other [14]. In English almost all the nouns can occur as verbs. But in Hindi verbalization of nominals is effected by combining two lexical items noun adjective adverb and a *simple* verb. For instance,

noun and verb	आरंभ करना	'arəmb^{h} kərna aarambha karanaa'	'to start'
adjective and verb	शांत करना	'ʃant kərna shaanta karanaa'	'to calm down'
adverb and verb	उठाकर रखना	'd^{h}ıre kərna uThaakara rakhanaa'	'to lift'.

According to traditional [15] and structural grammars [16], these verbs are classified as *conjunct verbs* with three sub-classes as shown above. From the syntax frames it is clear that the noun-verb combination is a true conjunct, as it gives a unique sense which is not decipherable from any other sources like sentence frames or semantic relations. On the other hand, the other two sub-classes can be deduced from sentence frames or through semantic relations. It is to be noted that the *compound verbs* in Hindi, *i.e*, a combination of a polar and a vector verb are dealt with in the manner of morphological processing. An instance of such verb is *गिर पड़ना, gir pəḍna, gira paRanaa* 'to fall down'.

In the Hindi WordNet, the *conjuct verbs* are stored through *conjuct-with* links between the first component (N Adv Adv) and the second (a simple verb). The verb hierarchy helps in optimizing the number of such links. The module for processing the compound verbs is a front end to the Hindi WordNet, just like the morphology module, and is table driven.

6 Results, Conclusions and Future Work

The work reported here started with English verbs. But these verbal concepts can be considered universal expressed using English alphabets. A hierarchy of English verbs has been created for the purpose of English Hindi machine translation. This hierarchy contains 5500 nodes (i.e. verbal concepts) corresponding to about 2000 unique English verbs. The principles behind organizing this hierarchy have been translated to Hindi, and a Hindi verb hierarchy too has been created. The top nodes in this hierarchy correspond to *act, move* and *put* classes in English. The verb hierarchy lends a structure to the organization of the verbs knowledge base in the Hindi WordNet. The coverage of both English and Hindi verbs is increasing everyday. A visualizer and an application programming interface for the verb knowledge bases in both the languages are under construction.

References

1. Jha Vishwanath, *Amarkosha by Amarsingha*, Motilal Banarasidas Publications, Varanasi, 1975.
2. Lenat D. B. and Guha R. V., *Building Large Knowledge Based System, Representation and Inference in the CYC Project.* Reading, Mass: Addison Wesley, 1990. `http://www.cyc.com`
3. `http://ontology.teknowledge.com/`
4. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. *Five Papers on WordNet.* CSL Report 43, Cognitive Science Laboratory, Princeton University, Princeton, 1990. `http://www.cogsci.princeton.edu/~wn`
5. Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.
6. Vossen Piek (ed.), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dodrecht. *Kluwer Academic Publishers,* 1998.
7. Chakrabarti Debasri, Narayan Dipak Kumar, Pandey Prabhakar, Bhattacharyya Pushpak, *Experiences in Building the Indo WordNet: A WordNet for Hindi.* Proceedings of the First Global WordNet Conference, 2002. `http://www.cfilt.iitb.ac.in/webhwn`
8. `http://framenet.icsi.berkeley.edu/~framenet`
9. `http://www.cis.upenn.edu/verbnet/`
10. Levin Beth, *English Verb Classes and Alternations A Preliminary Investigation.* The University of Chicago Press, 1993.
11. Dave Shachi and Bhattacharyya Pushpak, *Knowledge Extraction from Hindi Texts.* Journal of Institution of Electronic and Telecommunication Engineers, vol. 18, no. 4, July, 2001.
12. *The Universal Networking Language (UNL) Specifications*, Version 3.0, UNL center, UNDL Foundation, 2001. `http://www.unl.ias.edu/unlsys/unl/UNL%205specifications.html`.
13. Dowty, D., *Word Meaning and Montague Grammar*, Synthesis Language Library, Boston, 1979.
14. Dave Shachi, Parikh Jignashu and Bhattacharyya Pushpak, 2002, *Interlingua Based English Hindi Machine Translation and Language Divergence*, Journal of Machine Translation, Volume 17, September, 2002.
15. Bahari Hardev, *Vyavaharik Hindi Vyakaran Tatha Rachna.* Lokbharti Prakashan, Allahabad, India, 1997.
16. Singh Suraj Bhan, *Hindi ka Vakyatmak Vyakaran.* Sahitya Sahakar, Delhi, India, 1985.