Unification of Universal Word dictionaries Using WordNet Ontology and Similarity Measures

Sangharsh Boudhh, Pushpak Bhattacharyya

Center for Indian Language Technology (CFILT),
Department of Computer Science and Engineering,
Indian Institute of Technology (IIT Bombay), Mumbai – 76, INDIA
sboudhh@gmail.com, pb@cse.iitb.ac.in

Abstract. Interlingua-based Machine Translation systems work on parallel aligned lexicon of different languages. *Universal Networking language* uses *Universal Words* (*UWs*) as its lexicon. There exist some discrepancies in different language *UW dictionaries*. This poses a roadblock in interoperability of *UNL resources* created in different centers. We examine the challenge involved and develop a strategy to unify the dictionaries against the standard *U++ UW dictionary*. We exploit the WordNet ontology and closeness of *U++ UW dictionary* with it and the concept of similarity measures to recognize the semantically similar context.

Keywords: UNL, Universal Word, U++ UW dictionary, Hindi-UW dictionary, Extended Gloss Overlap, Lesk's algorithm, WordNet Ontology

1 Introduction

Interlingua-based Machine Translation systems require a standard unique lexicon and its linkage with different language words and attributes. Inter-lingual representation of a sentence contains only semantics represented using the standard lexicon. In order to translate this to any language L, we need a mapping from a lexeme entry of standard lexicon to language L words and corresponding attributes.

Attributes should consist of morphological, syntactic and semantic properties of the language L words. Standard interlingua lexicon should contain one or many disambiguated pivot entries of all concepts along with part of speech, their definition and usage examples. This lexicon does not contain any other attributes as it does not belong to any natural language.

The lexicons for language L is created at different places across the globe for different languages and they sometimes lack a standard to comply with while developing the lexicon. Our work describes an attempt at unifying such varied lexicons, in case of *Universal Words* (*UWs*) of *Universal Networking Language* (*UNL*).

1.1 Universal Networking Language (UNL)

UNL [6] is an Interlingua for representing information and knowledge provided by natural languages. A *UNL* representation is a hyper-graph with *Universal Words* or another *UNL* representation as a node. Nodes are connected with each other, wherever applicable, by links called *Relations*, which describe objective information of sentence. There are 41 relations defined at present. Universal words are further modified using *Attributes*, which describe subjective information such as speaker's point of view, time with respect to speaker, focus object, tense etc.

1.2 Universal Words (UWs)

Universal Words (*UWs*) are concept words, which form vocabulary of UNL. A UW consists of a *headword* (HW) which is a natural language word usually from English, followed by a set of restrictions which disambiguates UW to refer to only a specific sense of the HW.

Format. <UW> ::= <headword> [<constraint list>]

Example. *abbreviate(icl>reduce>do, agt>thing, obj>thing)*;
where abbreviate is the headword and rest is the constraint list. The keywords *icl*, *agt* and *obj* are taken from *UNL relations*. *agt>thing* implies agent is a "thing" or its subclass, similarly obj stands for object. *icl>reduce>do* implies this UW is a subclass of UW "reduce", which is consequently the subclass of "do", which is a verb.
The important *UNL relations* and their description are as follows:

Table 1. UNL relations and their description

relation	description
agt	agent – indicates a thing in focus that initiates an action
aoj	thing with attribute – indicates a thing that is in s state or has an attribute
equ	effected co-thing – indicates an equivalent concept
icl	included/a kind of – indicates an upper concept or a more general concept
iof	an instance of – indicates a class concept that an instance belongs to
obj	affected thing – indicates a thing in focus that is directly affected by an event or state

Categorization. UWs are hierarchically categorized. The most general concept is called *uw*, which is on top of the hierarchy. Then there are four categories i.e. nominal, verbal, adjectival, adverbial concepts represented by *thing*, *{do, occur, be}*, *adj*, *adv* respectively.

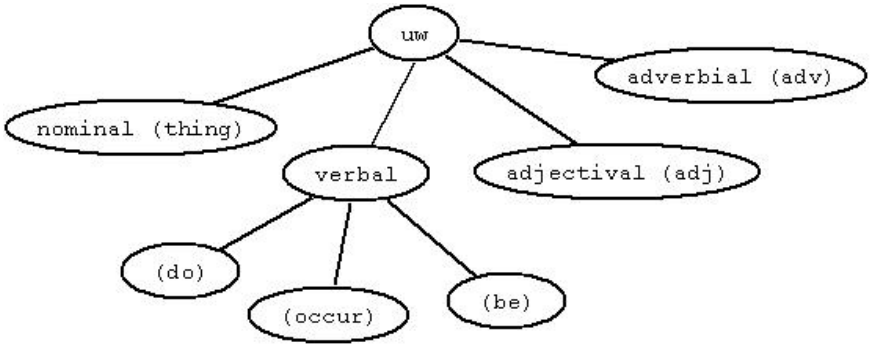


Figure 1. Universal word categorization

1.3 Unification problem

Different language centers have been following somewhat different guidelines for UW formation, due to unavailability of a standard *UW dictionary* till now. So the language words are linked with a set of *UWs* which is different from the UW dictionary released by *U++ Consortium*. This means on consulting UW dictionary of language L1 and L2 for same meaning word, we get two similar but different UWs. For the automatic translation to work, we must synchronize all these dictionaries with the standard one.

In most of the cases, we can assume headword to be the same, but UWs differ in constraints. We can see few entries from both the *U++* and *Hindi UW dictionary* for the headword dog. The U++ UW *dog(icl>canine>thing)* and Hindi-UW *dog(icl>animal)* represent the same concept but have been written differently. The similarity between these two is very evident to human. However for large lexicons, we need machines to be able to pick up the similarity with very high accuracy.

Table 2. UW entries for same word from different dictionaries

U++ UW dictionary	Hindi – UW dictionary
dog(icl>canine>thing)	dog(icl>animal)
dog(unpleasant_woman>thing, equ>frump)	dog(icl>constellation)
dog(icl>chap>thing)	dog(icl>mammal)
dog(icl>villain>thing, equ>cad)	dog(icl>female)

2 UW construction procedure

The guideline for formation of any new UW for some concept, as proposed in *U++ Consortium* meeting, July 2007 at Grenoble is described briefly [3]:

1. *Headword* Selection: Choose a word (HW) from English or some language, which completely covers the word W we are trying to describe
2. Ontological constraints:

- Noun: (*iof>X*), if W is instance of X and (icl>Y>thing), where Y is closest hypernym of W. e.g. *dog(icl>mammal>thing)*
- Verb: (*icl>do*) for action verbs, (*icl>occur*) for process describing verbs and (*icl>be*) for state denoting verbs.
- Adjective: (*icl>adj*)
- Adverb: (*icl>how*)

3. Semantic constraints: If HW is broader than W, restrict it using UNL relations (*rel>X*) and make equivalent to W.
   - (*icl>Z>Y*) for a narrower hypernym Z than Y. e.g. *dog(icl>canine>mammal)*
   - (*equ>S*) for synonym S. e.g.
   - (*ant>A*) for antonym A. e.g.
   - (*pof>A*), if W is part of A. e.g. room(pof>building)
   - (*icl<V*), for a hyponym V
4. Argument constraints: If W has some obligatory participants, which are usually present in sentence with W. e.g. agent or object; *give(agt>thing, obj>thing)*

3 UW dictionary

3.1 U++ UW dictionary

U++ UW dictionary [7] contains UW, part of speech information, definition and examples. The latest *U++ UW dictionary* has been derived from English WordNet [1][7] version 3.0 (EWN) and it can be backtracked to corresponding WordNet synset using sense key field. This is accepted as the standard dictionary by *U++ Consortium* members. It is maintained by Spanish language center.

Format. Format of *U++ UW dictionary* is:

UW;sense_key;pos_synset;freq_count

where first field, *UW*, is the Universal Word, *sense_key* is sense key of the corresponding entry in EWN 3.0, *pos_synset* is position of headword in the corresponding WordNet synset and *freq_count* is usage frequency for the corresponding synset. Using sense key, we can link the UW to a unique synset in EWN 3.0. A typical entry from U++ UW dictionary looks like:

dog(icl>canine>thing);dog%1:05:00::;0;42

3.2 L-UW dictionary (Hindi-UW)

Ideally, L-UW dictionaries should link the words of language L with U++ UWs and contain attributes for the words, examples in language L and other flags like

frequency. But the UWs used by many L-UW dictionaries are not same as U++ standard ones, as mentioned earlier.

The Hindi-UW dictionary [9] is made at Center for Indian Language Technology, IIT Bombay (India) under the supervision of Dr. Pushpak Bhattacharyya.

Format. Format of Hindi-UW dictionary is :

uniq_id; transliteration; hindi_stem; hindi_word; UW_headword; UW_restrictions; attributes; src_lang; priority; frequency; definition; example

Where *sense_key* is sense key of the corresponding entry in EWN 3.0, pos_synset is position of headword in the WordNet synset and freq_count is usage frequency for the synset. Using sense key, we can link the UW to a unique synset in EWN 3.0. A typical entry from *U++ UW dictionary* looks like:

saMkRipwa; xÉÇÍ¤ÉmiÉ; xÉÇÍ¤ÉmiÉ MüUlÉÉ; abbreviate; icl>reduce(agt>person,obj>thing) ; V,CJNCT,AJ-V,link,VOA,VOAACT, VLTN,TMP,obj-ko,Va; H; 0; 0; Abbreviate 'New York' and write 'NY'.; to shorten

1. Transliteration of Hindi stem - *saMkRipwa*
2. Hindi stem - *xÉÇÍ¤ÉmiÉ*
3. Hindi word - *xÉÇÍ¤ÉmiÉ MüUlÉÉ*
4. Headword of the UW - *abbreviate*
5. UW restrictions - icl>*reduce(agt>person,obj>thing)*
6. Attributes [9] - *V,CJNCT,AJ-V,link,VOA,VOA-ACT,VLTN,TMP,obj-ko, Va*
7. Source language(H for Hindi) - *H*
8. Frequency of usage - *0*
9. Priority of the word - *0*
10. Example - *Abbreviate 'New York' and write 'NY'.*
11. Explanatory meaning - *to shorten*

4 Observations and Statistics

Both the UW dictionaries U++ and IITB were examined for characteristics which could be helpful in the unification process.

4.1 Polysemy distribution

Distribution of number of entries per sense reflects the state of problem we would be facing. For *U++ dictionary*, we plot the number of senses a word can have within a part of speech. Here we consider the number of times the same headword appears under a part of speech. Since *IITB dictionary* is *L-UW dictionary* (*Hindi-UW*), here

we have to set the number of senses of an entry to be number of senses of the same headword, part of speech pair in *U++ UW dictionary*.

If number sense comes out to be zero in case of *IITB dictionary*, it means that there was no corresponding sense was found in *U++ dictionary*. If it is one, then there is single choice for alignment as there is only one sense possible for that word. For more than one senses, an algorithm has to be built to select the best possible.

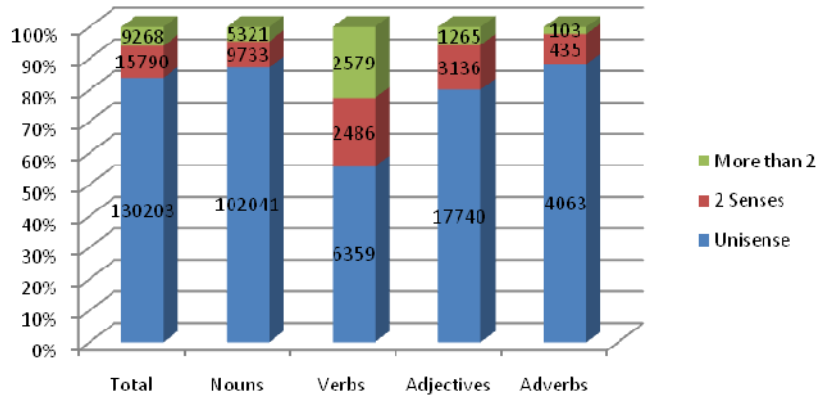


Figure 2. Distribution of senses in each PoS in U++ dictionary

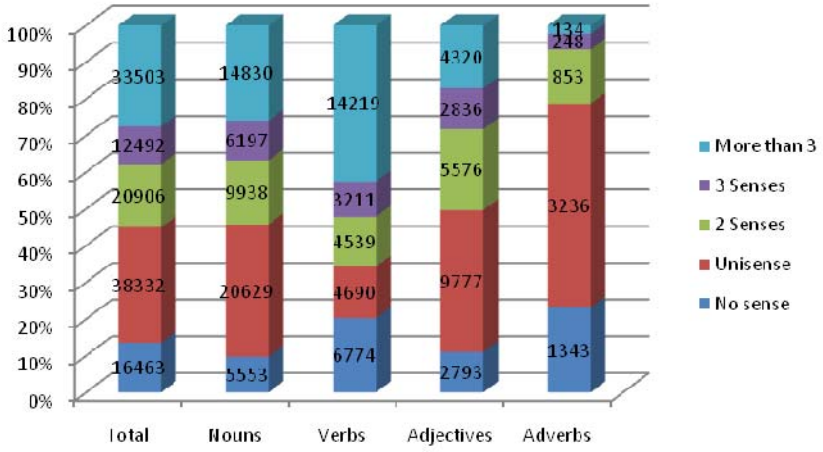


Figure 3. Distribution of senses in each PoS in IITB dictionary

4.2 Frequency of relations

We examined both the dictionaries and observed patterns in them, which can be exploited for their unification. We found out that out of 41 relations possible in UNL only a few appear in UWs and still very less is significant. Here is the percentage of UW in which the specified relation appears:

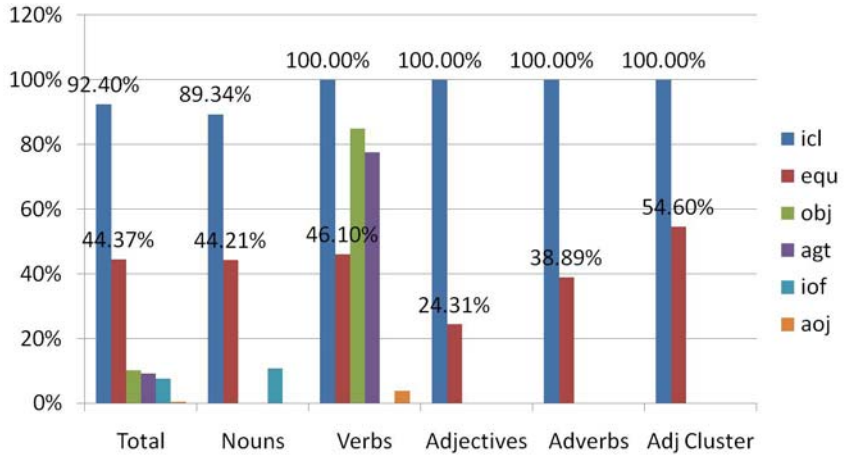


Figure 4. Frequency of occurrence of relations in different categories in U++ UW dictionary

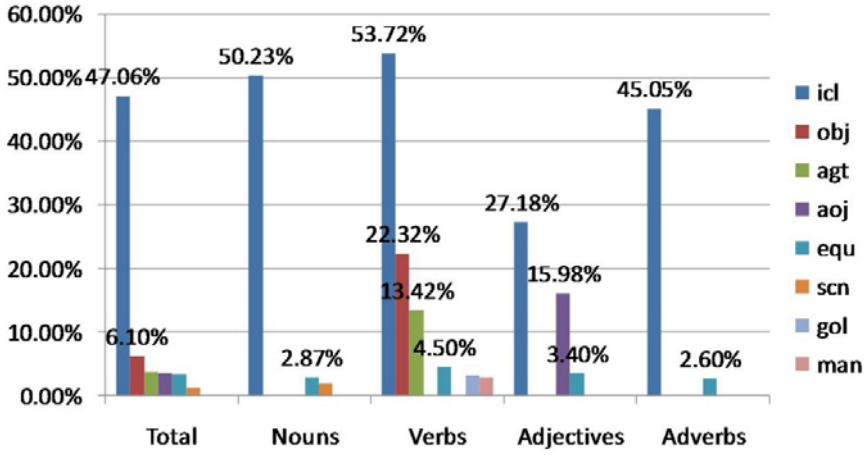


Figure 5. Frequency of occurrence of relations in different categories in Hindi-UW dictionary

As is evident from figure 2 and 3, *icl* is the most frequently used relation while defining UW. Other important relations are *agt*, *obj*, *equ*, *iof*, *aoj*. However the

frequencies are not same across the dictionaries because of the same fact of using different guideline in their formation.

4.3 Many to One relationship

All *U++ UWs* are linked to a synset of WordNet through *sensekey* and there is no duplicate *sensekey* in the dictionary. That means given a HeadWord and a Synset uniquely determines the U++ UW. However *IITB dictionary* is a *L-UW* one and can contains multiple entries for same UW, but with different Hindi words (synonyms). So if all the *IITB UWs* are mapped to *U++ UWs*, mapping will be many-one relationship.

4.4 "icl" and "equ" terms

"*icl*" term, in most cases, contains a direct (or indirect) hypernym of the UW. It is always in U++ dictionary and frequent in IITB. Similarly "*equ*" term in U++ dictionary belongs to first word of the same synset. However in IITB dictionary it may be a synonym or even hypernym in some cases.

5 Unification algorithm

5.1 Using Inter-lingua Index or Global WordNet grid

The following strategy works for the languages which have their WordNets aligned with the English WordNet. This is applicable for many European languages existing in Euro-WordNet e.g. Spanish, French, German etc. In order to create their UW dictionaries, synset from their WordNet can be mapped to corresponding synset in English WordNet, which in turns is linked with the standard UW dictionary. Here is the schematic diagram of the model [11] [12]:

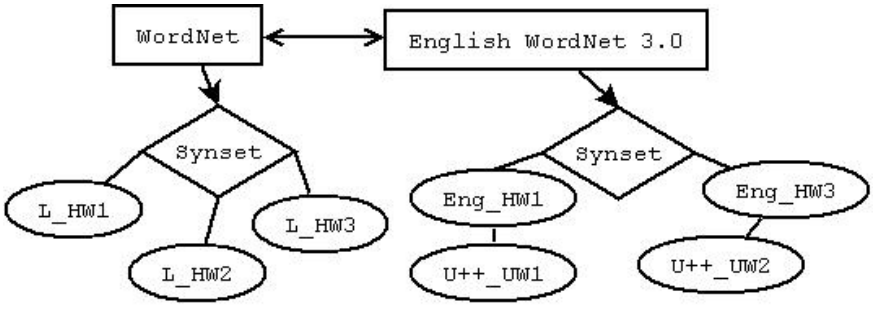


Figure 6. Unification using WordNet mapping

5.2 Using WordNet ontology and Similarity measures

Based on the observation, we developed an algorithm which is a combination of Ontology-based and Extended Gloss Overlap [2] algorithm.

1. Iterate through all the IITB UWs
2. Pick all candidate U++ UWs (those which share HeadWord and PoS)
3. Remove prefix ("the", "a", "be" etc.), if required.
4. Make a pair of *IITB UW* and candidate *U++ UWs*, one at a time
5. Calculate *SimpleMatch*, *RestrictionScore*, *GlossScore*, *ExampleScore*. Total score is sum of these four scores.
6. Declare the pair with maximum total score to be the aligned pair.
7. Impose a threshold score later to if the pair actually satisfies the minimum score criterion for acceptance.

5.2.1 Simple Match

This score is based on simple string matching for same relation terms, e.g. *icl-icl*, *iof-iof*, of *IITB* and *U++ UW* and *icl-equ*, *equ-icl* terms, matching of gloss pair, example pair after removing non-word characters and stop words. *icl-icl* means matching of the term with *icl* relation in *U++ UW* with the term with *icl* relation in *IITB UW*.

5.2.2 Restriction Score

For calculating restriction score, an inverted hypernymy tree is created keeping the *U++ UW synset* at the root and "*icl*", "*equ*" terms of IITB UW are searched in breadth first manner in the hypernymy tree. Score assigned is inversely proportional to the depth at which match is found.

5.2.3 Gloss and Example Score

All possible pairs of *IITB-U++ glosses* and *IITB-U++ examples* are considered. Firstly, non-word characters and stop words are removed. Then, maximal string overlap is calculated. Direct hypernym and hyponym glosses are also considered, inspired by Extended Gloss Overlap algorithm.

5.2.4 String Overlap Function

The string overlap function [13] breaks up the String in words and then further in letter pairs. For e.g. "like god" will be broken into "li", "ik", "ke", "go", "od". Then twice the number of common letter pairs is divided by total number of pairs.

For e.g. the score between "doing better" and "better do it" will be:

$$\frac{2 \times \left|(do, be, et, tt, te, er)\right|}{\left|(do, oi, in, ng, be, et, tt, te, er)\right| + \left|(be, et, tt, te, er, do, it)\right|}$$
$$= \frac{2 \times 6}{9 + 7} = 75\%$$

6 Results

Every aligned IITB UW has extra fields holding the *U++ UW* and its fields; number of candidate *U++ UWs* (same HW and PoS) and the four scores namely, *SimpleMatch*, *RestrictionScore*, *GlossScore* and *ExampleScore*.

6.1 Alignment Interface

A web interface for IITB to U++ UW alignment is also created. It provides an interface to search and select a IITB UW, look through its candidate U++ UW entries and select the one which user wants to align. Further it is supported with the scores to facilitate the user to make a better choice.

6.2 Results

Out of *124,862 UWs* in *IITB UW dictionary*, *3166* weren't considered as they don't belong to any of noun, verb, adjective or adverb part of speech. Out of remaining *121,696 UWs*, the algorithm has given score for *87287* entries. Following pie charts present the distribution of UWs with no sense found, not aligned UW, UW aligned with total Score greater than or equal to 50 for all parts of speech as well as in total. Here we are putting a criterion of total score to be greater than 50 for the calculation of precision.

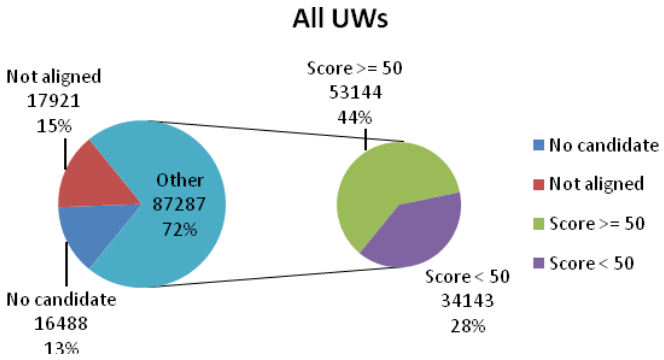
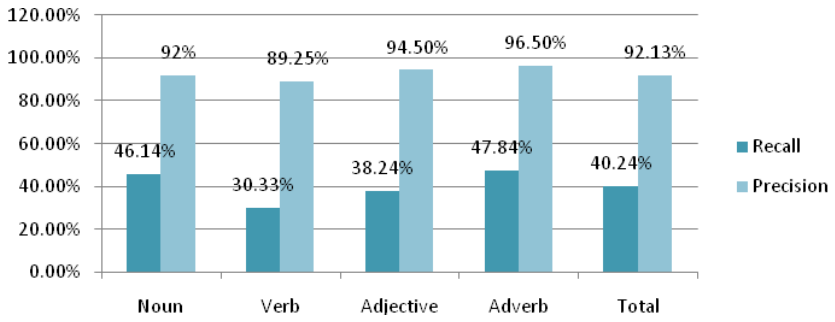


Figure 7. Distribution of alignment of IITB UWs

6.3 Recall and Precision

The alignments, with score greater than 50, are considered for recall and precision calculations. A set of 400 aligned UWs for each part of speech was randomly selected and manually checked to wrong alignments.

Part of Speech	Total number	Aligned(Score>=50)	Recall	Precision
Noun	57147	28662	46.14%	92%
Verb	33433	11361	30.33%	89.25%
Adjective	25302	10239	38.24%	94.5%
Adverb	5814	2882	47.84%	96.5%
Total	121696	53144	40.24%	92.13%



7 Analysis and Conclusion

7.1 Result Analysis

40.24% of UWs were aligned with precision of *92.13%* and further we have *28%* more UWs which are aligned with lesser precision. As we can see now, verbs among all were the toughest to align due to their highly polysemous behaviour and minute difference between two senses.

In case of unisense words, we don't have a choice of candidates to align with. But we have provided the scoring for the aligned pair in this case also, which enable us to know whether the alignment to unisense UW can be trusted or not.

Out of total *87287* aligned UWs, gloss functions gave score for *49246* entries, example functions for *44695* entries and restriction for *11937* entries. Although restriction is a very accurate way to establish alignment, its coverage is very less.

The *IITB UWs* which matched no sense in U++ dictionary mostly have multiword HeadWords, which are not so common occurrence in U++ dictionary. And the IITB UWs which had candidate but still weren't aligned are likely to be ones which lack two or three of restrictions, gloss and example and the fields present were not found to be "close" with candidate *U++ UWs* by the algorithm.

7.2 Conclusion

The exercise of aligning the IITB UW with U++ UW has various advantages. First of all, now it would be possible to deconvert UNL graphs created using standard U++ UWs at any place into Hindi with better quality output. And EnConverter of Hindi (when it comes) will also be able to create UNL graphs of globally accepted standard.

We may now merge the two UWs retaining some useful information from IITB UW also. There by enriching the U++ standard at the same time. Now since IITB UW is linked through sensekey to English WordNet also, it will get updated with each WordNet version also with more synonyms, better gloss and examples.

Although the algorithm has been created with IITB dictionary in mind, it can be easily extended to other L-UW dictionaries with similar scenario. As soon as all the countries adjust their system for U++ dictionary, the exchange of resources becomes easier and quicker. As far as we know, this is the first attempt at unification of a L-UW dictionary with U++ UW dictionary.

On the way of achieving this alignment, Java API for IITB dictionary and U++ dictionary were also created as by-products. Moreover, the interface created for manual alignment which shows scores from algorithm also will assist manual alignment to a great extent providing a graphical user interface and highlighting the more likely entities.

References

1. Christiane Fellbaum, WordNet: An Electronic Lexical Database, The MIT Press, 1998
2. Satanjeev Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pages 805–810, Acapulco, August.
3. Boguslavsky Igor, UW construction procedure, U++ Consortium meeting, Grenoble, July 2007
4. Pedersen, Patwardhan, and Michelizz: WordNet Similarity - Measuring the Relatedness of Concepts - Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), pp. 1024-1025, July 25-29, 2004, San Jose, CA (Intelligent Systems Demonstration)
5. Christian Boitet, Pushpak Bhattacharyya, Etienne Blanc1, Sanjay Meena, Sangharsh Boudhh, Vishal Vachhani , Georges Fafiotte, Achille Falaise Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a linguistic survival system under construction, SNLP, 2007
6. UNL specifications, http://www.undl.org/unlsys/unl/unl2005/
7. Universal Word dictionary, http://www.unl.fi.upm.es:8099/unlweb/
8. English WordNet, http://wordnet.princeton.edu/
9. Hindi-UW dictionary, http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php
10. Hindi WordNet documentation, http://www.cfilt.iitb.ac.in/wordnet/webhwn/
11. EuroWordNet, http://www.illc.uva.nl/EuroWordNet/
12. Global Wordnet grid: http://www.globalwordnet.org/
13. Overlap function: http://www.catalysoft.com/articles/StrikeAMatch.html