Unification of Universal Word dictionaries Using WordNet Ontology and Similarity Measures

Sangharsh Boudhh
Center for Indian Language Technology (CFILT)
IIT Bombay
sboudhh@gmail.com

Pushpak Bhattacharyya
Center for Indian Language Technology (CFILT)
IIT Bombay
pb@cse.iitb.ac.in

Abstract

We report an exercise that resembles ontology merging. Disambiguated words called *universal words (UW)* from two different sources are attempted to be unified through similarity computation. Using an Ontology based and Extended Gloss Overlap based algorithm, reasonable accuracy is obtained for nouns, followed by decreasing accuracy for adjectives, adverbs and verbs. The context is the Universal Networking Language (UNL) project which is an international endeavor for multilingual information access on the web.

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 disambiguated pivot entries of concepts along with their Parts of Speech, definition and usage instances. For example, the lexeme *spring* is ambiguous and has at least two meanings: *a season* and *a tool*. Possible unique meaning representations of these two senses are

spring(a-kind-of>season)
spring(a-kind-of>tool)

In absence of standardization, the same concepts can be expressed as

spring(a-kind-of>part-of-year)
spring(a-kind-of>instrument)

Humans typically do not have much problem dealing with such variations because of the large amount of world knowledge at their disposal (*season* is indeed a *part-of-year*; *tool* and *instrument* are synonyms). But automatic processes cannot operate correctly with such situations. To give an example, suppose it is required to translate between French and Hindi. To translate a sentence in French, meaning

Spring is a season of festivity

both the French analysis system and the Hindi generator system must agree that the *season* meaning of *spring* is involved. That is, both *French→Pivot* and *Pivot→Hindi* dictionaries should have uniform representation for this sense of *spring*.

The above discussion is in the context of an international project called the *Universal Networking Language (UNL)*[1] which was started in 1996 as an attempt to cross the language barrier on the web. 15 language groups from different parts of the world were involved in this endeavor. The idea was to encode (called *enconversion* in the *UNL* parlance*)* the sentences of a language L_1 into the UNL form and then generate (*deconversion* in the *UNL* parlance) the sentences of L_2 from the UNL form. It should be evident that both the languages must use the same pivot dictionary.

1.1 Universal Networking Language (UNL): the Framework

UNL is an electronic language for computers to express and exchange information (Uchida *et. al.* 2000). UNL expressions are generated sentence wise and consist of a set of directed binary relations, each between two concepts in the sentence. Tools called *EnConverter* and *DeConverter* which are language independent engines have been conventionally used for converting sentences from the source language to UNL and from UNL to the target language. The constituents of the UNL system are described now.

Universal Words

Universal words are the character-strings which represent simple or compound concepts. They form the vocabulary of UNL and represent the concepts in a sentence without any ambiguity. Universal Words may be simple or compound.

[1] *http://www.undl.org*

Simple unit concepts are called *simple UWs*. For example, *farmer(icl>person)* is a simple UW. Compound structures of binary relations grouped together are called Compound UWs. The syntax of a UW is given below.

<UW> ::= <Head Word> [<Constraint List>]
[<":"<UW-ID>] [“.”<Attribute List>]

where

(i) Head Word: is an English word interpreted as a label for a set of all the concepts that correspond to that word in English.

(ii) Constraint List: is the list of constraints that restricts the scope of the UW to a specific concept included within the Basic UW (explained next).

(iii) UW-ID: is an identifier used to indicate some referential information.

Attributes

Attributes of Universal Words describe the subjectivity of the sentence. They provide information about how a concept is used in a given sentence. The attributes enrich the information content of the UNL by providing information like logicality of UW, time with respect to the speaker, speaker's view on aspects of the event, speaker's view of reference to the concept, speaker's view on emphasis, focus and topic, speaker's attitudes, and speaker's feelings and judgments. The UNL group has provided a very rich set of attributes which makes it possible to capture many real world situations in the UNL form. Currently, there are 87 attribute labels. Some of the attributes are: @past, @present, @future, @imperative, @interrogative, @passive, @topic, @intention, *etc*.

UNL Relations

Binary relations of the UNL expressions represent directed binary relations between the concepts of a sentence. There are a total of 46 relation labels defined in the UNL specifications.

We classify the semantic relations (with overlapping) as the following:

a. Relations between two entities <e_1, e_2>, where e_1 is a verbal concept (29 relations)

b. Relations between two entities <e_1, e_2>, where e_1 is a non-verbal concept

	Arguments <e_2>	Adjuncts <e_2>
DO <e_1>	agt bas ben cag cob con coo dur gol ins obj opl ptn pur rsn scn seq src	man met plc plf plt via tim tmf tmt
Occur <e_1>	ben cob con coo gol obj opl rsn scn seq src	dur man plc plf plt via tim tmf tmt
BE <e_1>	aoj bas ben cao cob con coo dur gol obj plc rsn scn src	plf plt tim tmf tmt man

Table 1: Relations for Verbal Concept

UNL Graph

The UNL representation of a sentence is expressed in the form of a semantic graph, called *UNL graph*. Consider the sentence (1).

(1) *John eats rice with a spoon.*

The UNL expression for (1) is given in (2) and the the UNL graph is illustrated in Figure 1.

(2) [UNL:1]
agt(eat(icl>do).@entry.@present, John(iof>person))
obj(eat(icl>do).@entry.@present, rice(icl>food))
ins(eat(icl>do).@entry.@present, spoon(icl>artifact))
[\UNL]

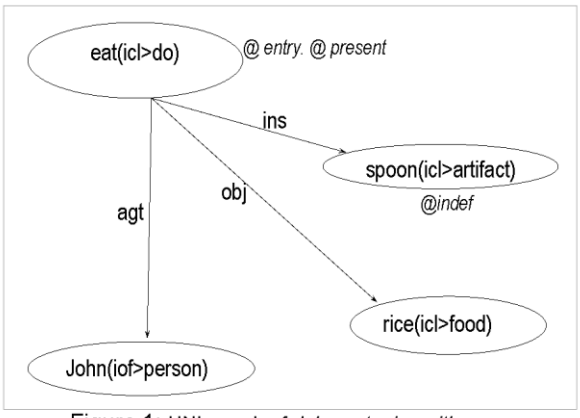

Figure 1: UNL graph of *John eats rice with a spoon*

In figure 1, the arcs are labeled with *agt* (agent), *obj* (object) and *ins* (instrument), and these are the semantic relations in UNL. 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 for the purpose of denoting unique sense. *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* (*used for abbreviations*)- can be attached to an UW for restricting its sense.

UNL Hypergraph

UNL has a way of representing coherent sentence parts (like clauses and phrases). It uses the notation *:0<n>* where *<n>* is an integer. Compound UW (also called a scope node) is like a graph within a graph and has its own entry node. Compound UWs are powerful constructs in UNL. Scope is a mechanism used in the UNL format to express compound concepts in a sentence as well as coordinating concepts. Clauses can be considered as compound concepts and these are usually marked with a scope. For example, the UNL expression, omitting the UNL restriction information, for the sentence (3) is given in (4).

(3) Mary claimed that she had composed a poem.
(4) [UNL:3]
agt(claim.@entry.@past, Mary)
obj(claim.@entry. past, :01)
agt:01(compose.@past.@entry.@complete, she)
obj:01(compose.@past.@entry.@complete,poem.@indef)
[\UNL]

The segment *she had composed a poem* is considered as being within a scope, with the predicate *compose* being the entry node. The entire scope is connected to the matrix verb *claim* through the *obj* relation. The scope is represented in the UNL expression by the compound UW ID :01. Any compound concept can be represented using a scope and the scope technique allows us to capture deeply nested constructs in the language.

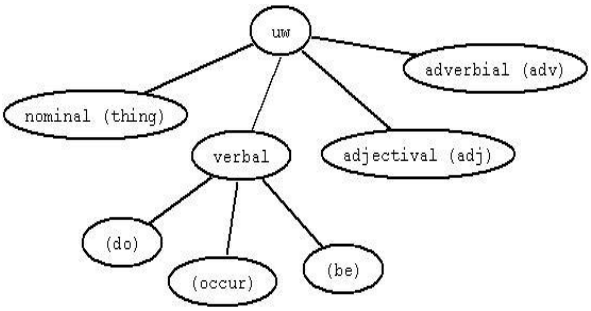

Figure 2: Universal word categorization

Categorization of UWs

UWs are hierarchically categorized (figure 2 above). The most general concept is called *uw*, which is on top of the hierarchy. Then there are four categories, *viz.*, nominal, verbal, adjectival, adverbial concepts represented by *thing*, *{do, occur, be}*, *adj*, *adv* respectively.

2 Problem Definition

Different language groups at different places of the world have been following somewhat different guidelines for UW formation, resulting in non-standard representation for UWs. We have studied two biggest UW dictionaries currently existing in the UNL community. The first is the dictionary developed at Madrid by the U++ consortium[2]: we will call this the **U++ dictionary**[3]. The second is the Hindi-UW dictionary[4] developed for the purpose of enconversion from and deconversion into Hindi at the Center for Indian Language Technology (CFILT [5]), Computer Science and Engineering department, Indian Institute of Technology Bombay (IIT Bombay[6]). We will call this the **H-UW dictionary**.

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)

Table 2: UW entries for same word from different dictionaries

We can see a few entries from both the *U++* and *Hindi-UW dictionary* for the headword *dog* (Table 2). The U++ UW *dog(icl>canine>thing)* and the H-UW *dog(icl>animal)* represent the same concept but have been written differently.

The similarity between these two is very evident to human. However it is non-trivial for a

[2] U++ Consortium is an open and free association of researchers, business entities and people with a common interest in the development of useful applications to society based in the UNL language. Its main interest focuses on the creation of applications to support multilinguality, in order to overcome linguistic barriers on the Internet.
http://www.unl.fi.upm.es/consorcio/index.php

[3] U++ UW dictionary (web interface)
http://www.unl.fi.upm.es:8099/unlweb

[4] http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php

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

[6] http://www.iitb.ac.in

machine to be able to pick up this similarity with very high accuracy.

Our goal in this work is to unify the two dictionaries mentioned above, *i.e.*, create mappings between UW entries from the U++ dictionary and the H-UW dictionary.

3 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 [Boguslavsky 2007]:

1. Headword Selection: Choose a word (HW) from English or some language (but using Roman Alphabet), 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 the HW is broader in scope 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)*

4 The two UW dictionaries

The structures of two types of UW dictionaries i.e. U++ UW dictionary and L-UW dictionary have been explained in this section. L-UW dictionary is explained with an example of H-UW dictionary which has language L as Hindi which is also more relevant for our context.

4.1 U++ UW dictionary

U++ UW dictionary contains UW, part of speech information, definition and examples. The latest *U++ UW dictionary* has been derived from English WordNet[7] [Fellbaum 1998] version 3.0 (EWN) and entries in it can be traced to corresponding WordNet synsets using the sense key field. This is accepted as the standard dictionary by *U++ Consortium* members. It is maintained by the Spanish language center.

The format of *U++ UW dictionary* is:

UW; sense_key; pos_synset; freq_count

where the first field *UW*, is the Universal Word, *sense_key* is the sense key of the corresponding entry in EWN 3.0, *pos_synset* is the position of headword in the corresponding WordNet synset and *freq_count* is usage frequency for the corresponding synset. Using the 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

4.2 H-UW dictionary

The H-UW dictionary[8] is made at the Center for Indian Language Technology[9], IIT Bombay (India) under the supervision of Dr. Pushpak Bhattacharyya.

The format of Hindi-UW dictionary is :

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

A typical entry from the H-UW dictionary looks like:

saMkRipwa; संक्षिप्त; संक्षिप्त करना; 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 - संक्षिप्त
3. Hindi word - संक्षिप्त करना
4. Headword of the UW- abbreviate
5. UW restrictions - icl>reduce(agt>person,obj>thing)

[7] http://wordnet.princeton.edu/

[8] Hindi-UW dictionary: http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php

[9] CFILT: http://www.cfilt.iitb.ac.in/

6. Attributes- 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

5 Some observations on the U++ and Hindi-UW dictionaries

Statistical data gathered from U++ and H-UW dictionaries and inferences derived from them have been explained in following sub-sections.

5.1 Polysemy distribution

Distribution of number of entries per sense reflects the complexity of problem we would be facing. Tables 3 and 4 show the polysemy distribution in U++ and H-UW dictionaries respectively.

	Unisense	2 Senses	More than 2
Total	130203	15790	9268
Nouns	102041	9733	5321
Verbs	6359	2486	2579
Adjectives	17740	3136	1265
Adverbs	4063	435	103

Table 1: Distribution of senses in each PoS in U++ dictionary

	Unisense	2 Senses	More than 2
Total	166463	38332	20906
Nouns	5553	20629	9938
Verbs	6774	4690	4536
Adjectives	2793	9777	5576
Adverbs	1343	3236	853

Table 2: Distribution of senses in each PoS in H-UW dictionary

5.2 Frequency of relations

We found that out of the 46 semantic relations in UNL only a few appear in the UWs. Here is the percentage of UW in which specific relations appear:

	Total	Nouns	Verbs	Adjs	Advs
icl	92.4%	89.3%	100%	100%	100%
equ	44.3%	44.2%	46.1%	24.3%	38.9%
obj	10.1%	0	85%	0	0
agt	9.21%	0	77.5%	0	0
iof	7.6%	10.7%	0	0	0

Table 3: Frequency of occurrence of relations in different categories in U++ UW dictionary

	Total	Nouns	Verbs	Adjs	Advs
icl	47%	50.2%	53.7%	27.2%	45%
equ	3.4%	2.9%	4.5%	3.4%	2.6%
obj	6.1%	0	22.32%	0	0
agt	3.7%	0	13.4%	0	0
aoj	3.4%	0	0	16%	0

Table 4: Frequency of occurrence of relations in different categories in Hindi-UW dictionary

As is evident from tables 5 and 6, *icl* (meaning *a-kind-of*) is the most frequently used relation while defining UW. Other important relations are *agt* (agent), *obj* (object), *equ* (synonym), *iof* (instance-of), *aoj* (attribute-of-object).

6 Unification algorithm

We developed an algorithm which is a combination of Ontology based and Extended Gloss Overlap based [Banerjee and Perdersen 2003; Perdersen *et. al.* 2004] algorithms:

Basic Pseudo Code

```
foreach UW U in L-UW dictionary {
  upp_uws[] = All U++ UWs with sameHeadWord and Part of Speech as U;

  pairs[] = U and elements of upp_uws one by one;

  foreach element in pairs[]{
        TotalScore = SimpleMatch() + RestrictionScore() + GlossScore() + ExampleScore();
        }

  best_pair = Element from pairs[] with maximum TotalScore;

  if(best_pair.TotalScore >= THRESHOLD_SCORE){
        Finalize and store that pair;
        }
}
```

Simple Match

This score is based on simple string matching for same relation terms, e.g. *icl-icl*, *iof-iof*, of *H-UW* and *U++ UW* and *icl-equ*, *equ-icl* terms, match-

ing 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 *H-UW UW*.

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 H-UW UW are searched in breadth first manner in the hypernymy tree. The score assigned is inversely proportional to the depth at which match is found.

Gloss and Example Score

All possible pairs of H-UW and U++ glosses and H-UW and 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.

String Overlap Function

The string overlap function[10] breaks up the string into words and then further into letter pairs. For example, *like god* will be broken into "li", "ik", "ke", "go", "od". Then two times the number of common letter pairs is divided by the total number of pairs.

For example, the score between "doing better" and "better do it" will be:

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

$$=\frac{2\times 6}{9+7}=75\%$$

7 Results

Out of the 121696 noun-adjective-verb-adverb *UWs* in H-UW dictionary, our algorithm could score 87287 entries.

Status of H-UW UW	Count	Percentage
No. of candidates	16488	13%
Not aligned	17921	15%
Score >= 50	53144	44%
Score < 50	34143	28%

Table 5: Distribution of alignment of H-UW UWs

Table 7 shows the distribution of UWs with *no sense found, not aligned* UW, UW aligned with total *score greater than or equal to 50* as well as those with *score less than 50*.

Recall and Precision

The alignments, with score greater than 50, are considered for recall and precision calculations.

PoS	Total number	(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%

Table 6: Recall and Precision for all Parts of Speech for a threshold score of 50.

8 Results and Discussions

40.24% of UWs were aligned with a precision of 92.13%. Verbs were the toughest to align due to their highly polysemous behaviour and minute difference between senses. Out of the 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 small.

UWs in H-UW dictionary which matched no sense in the U++ dictionary mostly have multi-word HeadWords.

9 Conclusion and future work

The exercise of aligning the H-UW 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

[10] http://www.catalysoft.com/articles/StrikeAMatch.html

be able to create UNL graphs of globally accepted standard.

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

Related pieces of work are by Ponzetto and Navigli (2009) and Ehrig and Sure (2004). The latter proposes to use category theory to provide a scheme independent ontology mapping, while the former concentrates on WordNet and Wikipedia mapping.

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

Future work is directed at improving the recall of the alignment.

References

Christiane Fellbaum (*ed.*). 1998. *WordNet: An Electronic Lexical Database*, The MIT Press.

Hiroshi Uchida, M. Zhu, and T. Della. Senta. 1999. *UNL: A Gift for a Millennium.*The United Nations University, Tokyo.

Igor Boguslavsky. 2007. *UW construction procedure*, notes of U++ Consortium meeting, Grenoble.

M. Ehrig and Y. Sure. 2004. *Ontology mapping – an integrated approach*. In Bussler C., Davis J., Fensel D. and Studer, R., eds., Proceedings of the First European Semantic Web Symposium, volume 3053 of Lecture Notes in Computer Science. Heraklion, Greece.

Satanjeev Banerjee and Ted Pedersen. 2003. *Extended gloss overlaps as a measure of semantic relatedness*. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI).

Simone Paolo Ponzetto, Roberto Navigli. 2009. *Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia*. International Joint Conference on AI (IJCAI).

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. *Wordnet::Similarity - Measuring the Relatedness of Concepts*. In Daniel Marcu Susan Dumais and Salim Roukos, editors, HLT-NAACL 2004: Demonstration Papers, pages 38--41, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.