# Lexical Resources for Semantic Extraction

**Rajat Kumar Mohanty, Pushpak Bhattacharyya**
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai – 400076, INDIA
Email: {rkm, pb}@cse.iitb.ac.in

**Abstract**

In this paper, we report our work on the creation of a number of lexical resources that are crucial for an interlingua based MT from English to other languages. These lexical resources are in the form of sub-categorization frames, verb knowledge bases and rule templates for establishing semantic relations and speech act like attributes. We have created these resources over a long period of time from Oxford Advanced Learners' Dictionary (OALD) [1], VerbNet [2], Princeton WordNet 2.1 [3], LCS database [4], Penn Tree Bank [5], and XTAG lexicon [6]. On the challenging problem of generating interlingua from domain and structure unrestricted English sentences, we are able to demonstrate that the use of these lexical resources makes a difference in terms of accuracy figures.

## 1. Introduction

Recent times are witnessing a revival of interest in and realizing the importance of rule and knowledge based methods in NLP - particularly MT. In statistical MT, the notion of language resources is confined to parallel corpora; but one quickly sees the saturation of the accuracy figure beyond a limit in statistical approaches. It is being increasingly felt that the involvement of human constructed lexical resources is inevitable in ensuring high levels of accuracy. We report here our work on the creation of a number of lexical resources that are crucial for an interlingua based MT from English to other languages.. These lexical resources are in the form of sub-categorization frames, verb knowledge bases and rule templates for establishing semantic relations and speech act like attributes. We have created these resources over a long period of time from Oxford Advanced Learners' Dictionary (OALD) (Hornby, 2001), VerbNet (Schuler, 2005), Princeton WordNet 2.1 (Miller, 2005), LCS database (Dorr, 1993), Penn Tree Bank (LDC, 1995), and XTAG lexicon (XTAG Research Group, 2001). On the challenging problem of generating interlingua from domain and structure unrestricted English sentences, we are able to demonstrate that the use of these lexical resources makes a difference in terms of accuracy figures.

## 2. Universal Networking Language: the framework

UNL is an electronic language for computers to express and exchange information (Uchida et.al., 1999). The three building blocks of UNL are (i) Semantic Relations, (ii) Attributes and (iii) Universal Words. The UNL representation of a sentence is expressed in the form of a semantic net (Woods, 1975) 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 UNL graph is 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]

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 disambiguating restrictions in parentheses. *icl* stands for *inclusion* and *iof* 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. Of special significance is the *@entry* attribute. This is typically attached to the main predicate.

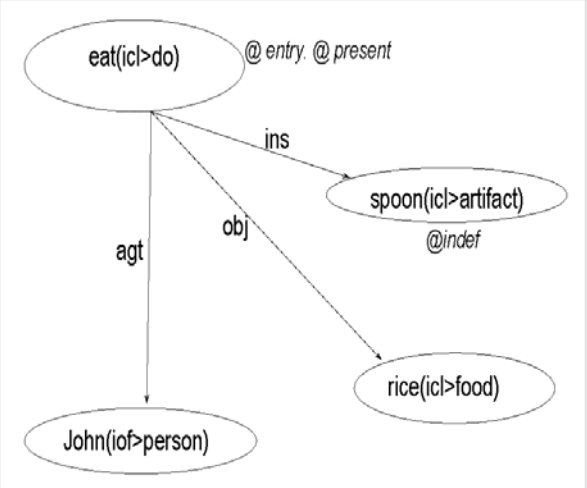


Figure. 1: UNL graph of *John eats rice with a spoon*

### 2.1 UNL Hypergraphs: a way of representing embeddings

UNL has a way of representing coherent sentence parts (like clauses and phrases) through *Compound UWs* also called *scope nodes*. These scope nodes are like graphs within graphs. These subgraphs have their own environment and the *@entry* node. For example, the UNL expression 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 clause *she had composed a poem* is considered as being within a scope, with the predicate *compose* being the entry node. The entire compound word or scope- as it is called- is connected to the matrix verb *claim* through the *obj* relation. Note that the scope is given a compound UW ID :01 to denote a separate environment.

### 2.2 SRS: a step towards UNL generation

A Semantically Relatable Sequence (SRS) (Mohanty et. al., 2005) of a sentence is defined to be a group of unordered words in the sentence (not necessarily consecutive) that appear in the semantic graph of the sentence as linked nodes or nodes with speech act labels, as illustrated in (5).

(5) John spoke to the students.
[SRS:5]
(John, spoke.@entry )---------- (CW, CW)
(spoke,@entry, to, students)----- (CW, FW, CW)
(the, students)---------- (FW, CW)
[\SRS]

*Once a sentence is broken into SRSs, no structural ambiguity remains to be resolved.* Each SRS need to be converted to a semantic relation with arguments or is translated into the UNL attribute labels, as illustrated in (6) for the SRSs given in (5).

(6) [UNL:5]
agt (speak.@entry.@past, John)
gol(speak.@entry.@past, student.@pl.@def)
[\UNL]

In the subsequent sections, we show the creation of knowledgebase using a number of existing lexical resources which is, in turn, used in generation of UNL expressions.

## 3. Knowledge Bases (KB)

The knowledgebase used for UNL generation consists of a Subcategorization Knowledge Base, a Verb Knowledge Base, a Lexical Knowledgebase with semantic attributes leading to the UNL Relation Rule Base, and a database of functional elements with grammatical attributes leading to UNL Attribute Rule Base. These knowledge bases are used at different stages of the sentence processing starting from parse tree correction to UNL attribute generation. Figure 2 gives an overview of the different knowledgebase and the associated lexical items which contribute towards SRS and UNL generation.

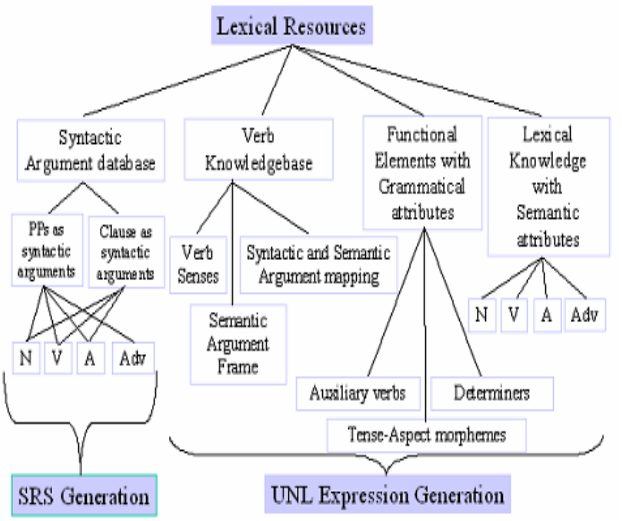


Figure. 2: An overview of our lexical resources

### 3.1 Lexical Subcategorization Knowledge Base

A lexical item, whether it is a noun, a verb, an adjective or an adverb, requires certain elements in order to appear in a meaningful sentence. For instance, the verb *put* takes a locative PP as its argument to complete its syntactic requirement in a sentence. The expression *[John put a book]* is considered to be an incomplete sentence, whereas *[John put a book on the table]* is correct. The locative PP *on the table* is said to satisfy the subcategorization requirements of the verb *put*. Lexical items subcategorize noun phrases (NP), prepositional phrases (PP) and clauses of different kinds to express their meaning in a sentence (Chomsky, 1981). This syntactic information is crucial at various stages of in-depth sentence analysis or meaningful sentence generation. We term these subcategorized phrases/ clauses as *syntactic arguments.*

The necessary information on these subcategorized elements is obtained from the Oxford Advanced Learners' Dictionary (OALD) (Hornby, 2001) manually. For instance, a verb like *donate* subcategorizes one NP and one PP. The information that the PP must be a *to-PP* is extracted from the notation *~ sth (to sb/sth)* given in the entry of *donate* in the OALD. Similarly, the entry of the adjective *jealous* subcategorizes an *of-PP* which is extracted from the notation *~(of sb/sth)* given in the OALD. Currently, the coverage of lexical entries in our subcategorization knowledge base is same as the coverage of WordNet 2.1 (Miller, 2005).

### 3.2 Verb Knowledge Base (VKB)

The motive behind the Verb Knowledge Base (VKB) creation is to provide all the necessary information of

verbs (*lexical, syntactic* and *semantic*) under one umbrella. It is being developed using various linguistic insights and many existing resources like OALD (Hornby, 2001), VerbNet (Schuler, 2005), WordNet 2.0 (Miller, 2005), Lexical Conceptual Structure (LCS) database (Dorr, 1993). None of the existing lexical resources is comprehensive enough to provide all these information together, though each one is specific in some way. For instance, WordNet provides rich lexical semantic information, but poor syntactic information associated with the lexical items. The OALD is rich enough to provide detailed subcategorization information, but poor in lexical semantics information.

### 3.2.1 The Inadequacy of VerbNet

The information obtained from VerbNet (Schuler, 2005) is not enough for semantics extraction for the following reasons: the semantic fields of verbs are not specified appropriately following the *Thematic Relations Hypothesis (TRH)*[1] (Gruber, 1965). On the other hand, in our Verb Knowledge Base, we adapt the *TRH* (Gruber, 1965) and the subsequent work ((Dorr, 1993), (Jackendoff, 1990)) to provide the basic conceptual functions of verbs in terms of *BE*, *DO* and *OCCUR* along the lines of $[_{State}$ *BE]*, $[_{Event}$ *GO]* and $[_{Event}$ *STAY]*. We use the WordNet to obtain the disambiguating restrictions of lexical items in terms of Universal Words (UNDL Foundation, 2006), the OALD to obtain the syntactic argument of verbs, the LCS database (Dorr, 1993) and the VerbNet (Schuler, 2005) to obtain the thematic role information of verbs.

Table 1 illustrates the structure of the Verb Knowledge Base entry. We obtain the initial set of semantic argument frames (described in Table 3) using the LCS database (Dorr, 1993) , WordNet 1.6, WordNet 2.0, the mapping database of WordNet 1.6 and WordNet 2.0. Later we manually enrich the knowledgebase. At present, the verb knowledge base contains about 22,000 entries.

| Base Verb | Disambiguating restrictions of Lexical items of Universal words: (icl>UW) | Basic Conceptual Function: BE \| DO \| OCCUR | [Semantic Argument Frame] | SyntacticArgument-SemanticArgument mapping: REL: ARG –TYPE | Example sentence |
|---|---|---|---|---|---|
| transfer | (icl>cause to change ownership) | DO | [agt>thing, obj>thing, gol>thing] | gol:to | I transferred my stock holdings to my children. |
| transfer | (icl>move) | DO | [agt>thing, obj>thing, src>thing, gol>thing] | gol:to src:from | He transferred the packet fromhis trouser pocket to a pocket in his jacket. |
| sleep | (icl>rest) | BE | [aoj>thing] | - | He sleeps in the morning. |
| crack | (icl>break) | OCCUR | [obj>thing] | - | The glass cracked when it was heated. |
| crack | (icl>hit) | DO | [agt>thing, obj>thing] | - | The teacher cracked him hard. |

Table 1: Verb Knowledge Base (VKB) Structure

[1] The *Thematic Relation Hypothesis (TRH)* (Gruber, 1965) states that the semantic fields differ in three possible ways: (i) what sorts of entities may appear as theme, (ii) what sorts of entities may appear as reference objects, (iii) what kind of relation assumes the role played by location in the field of spatial expressions.

## 3.3 Lexical Knowledge with Semantic Attributes

The Lexical Knowledge includes the semantic attributes of lexical elements (*i.e., Noun, Verb, Adjective* and *Adverbs*) which are represented through the UNL Relation Rule Base. The Relation Rule Base is one major component of the UNL generation system. For each rule- responsible for generating a semantic relation- a specific relation generation template is used. The OALD (Hornby, 2001), VerbNet (Schuler, 2005), WordNet 2.1 (Miller, 2005), and Treebank (LDC, 1995) are exploited for this. The number of rules is 741 at present, and is getting enriched everyday. The rule template is depicted in Table 2 and is explained in the subsections below.

| CW1 | | | | FW | | | | CW2 | | | | REL(UW1,UW2) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Syntactic Feature | | Semantic Feature | | Syntactic Feature | | Semantic Feature | | Syntactic Feature | | Semantic Feature | | | |
| SynCat | POS | SemCat | Lex | SynCat | POS | SemCat | Lex | SynCat | POS | SemCat | Lex | Rel | (UW1,UW2) |
| - | - | V020 | - | - | - | - | into | N | - | - | - | gol | (CW1,CW2) |
| V | - | - | - | - | - | - | within | N | - | TIME | - | dur | (CW1,CW2) |

Table 2: Rule Template

### 3.3.1 Syntactic Features

The field Syntactic Feature in the rule template consists of two subfields: syntactic category (SynCat) and Parts-Of-Speech (POS). The SynCat field is defined to be one of the broad lexical categories, such as N, V, J, R and P corresponding to nouns, verbs, adjectives, adverbs, and prepositions, respectively. The POS field is filled with the parser generated POS tags which are specific to the particular inflected items.

### 3.3.2 Semantic Features

This field consists of two subfields: semantic category (SemCat) and the actual lexical item (Lex). The Lex field is filled only when it is very specific as in the case of FW or when the SemCat field is not yet defined. As of now, the SemCat field is defined for verbs, nouns, and adverbs.

**Verbs:** The SemCat field for verbs carries the semantic grouping of verbs on the basis of our analysis on OALD data and Levin's Verb classification (Levin, 1993) from VerbNet data (Schuler, 2005). Each verb group is stored in a table, and is mapped to the SemCat field in terms of a unique ID. For example, the ID *v115* in the SemCat field is for *Contribute Verbs*, while *v139* is for *Meet Verbs* and *vErg* is for *Ergative Verbs.*

There are 189 verb groups for 4115 unique verbs. The rules using one of these 189 classes are developed from VerbNet and OALD data. The relevance of such a rule in UNL relation generation for the SRS *(cut, with, knife)* is illustrated in Table 3.

| CW1 | FW | CW2 | UNL Relation Generated |
|---|---|---|---|
| V _ v139 _ | P IN _ with | N _ _ _ | ins |
| V _ v024 _ | P IN _ to | N _ _ _ | gol |

Table 3: Illustration of Rules having semantic attributes of verbs in terms of verb groups (e.g., v139, v024, etc.)

The last column in Table 3 is an *action*, while the previous columns are *conditions*. The whole row stands for the rule that

*if*

there is an SRS of the form *($CW_1$, FW, $CW_2$)*, where *$CW_1$ is a verb in the 139$^{th}$ verb class* and *FW is the preposition 'with'* and *$CW_2$ is a noun*

*then*

*insert the relation 'ins' (instrument) between $CW_1$ and $CW_2$*

**Nouns:** The SemCat field for nouns carries the semantic grouping of nouns on the basis of the WordNet 2.1 (Miller, 2005) noun classification. Semantic features like TIME, PLACE, ANIMATE, INSTRUMENT, LEGAL DOCUMENT *etc.* are detected using the hypernymy hierarchy of words in the WordNet.

**Adverbs**: The SemCat field for the adverbs carries the semantic grouping of adverbs on the basis of the classification done in the Penn Tree Bank Release II (LDC, 1995). The lexical items having tags like ADV-MNR (adverb of manner), ADV-TMP (adverb of time) and ADV-LOC (adverb of location) are acquired from the Treebank, and are encoded in the Rule Base. Table 4 illustrates three rules for the SRSs containing adverbs, generating, *man (manner), tim (time)* and *plc (place)* relations- example situations being *(playing, well), (coming, early), (playing, there).*

| CW1 | FW | CW2 | UNL Relation Generated |
|---|---|---|---|
| V _ _ _ | _ _ _ _ | R _ MNR _ | man |
| V _ _ _ | _ _ _ _ | R _ TMP _ | tim |
| V _ _ _ | _ _ _ _ | R _ LOC _ | plc |

Table 4: Illustration of Rules having semantic attributes of Adverbs

### 3.4 Functional Elements with Grammatical Attributes

The grammatical information expressed by the functional elements is represented through the UNL Attribute Rule Base. For example, the attribute *@passive* is generated from *(<be-aux>, VBN)* type SRSs found in a sentence like "*This letter must have been written by her*". There are different combinations of modals, auxiliaries and verb-forms- VBD, VBN, VBZ, *etc.* obtained from the parser output- which are used to create the UNL attributes for verbs. Similarly, there are rules to generate attributes for nouns and adjectives.

| String of FWs | CW | UNL attribute list generated |
|---|---|---|
| has_been | VBG | .@present.@complete.@progress |
| has_been | VBN | .@present.@complete.@passive |
| should_ have_been | VBN | .@past.@complete.@obligation. @passive |

Table 5: Attribute Generation Rule Base Template

Table 5 illustrates this. As usual, the first three columns give the conditions for generating the attribute in a particular row.

## 4. Overview of the System

The system architecture is shown in figure 3, in which the knowledge base is used heavily.

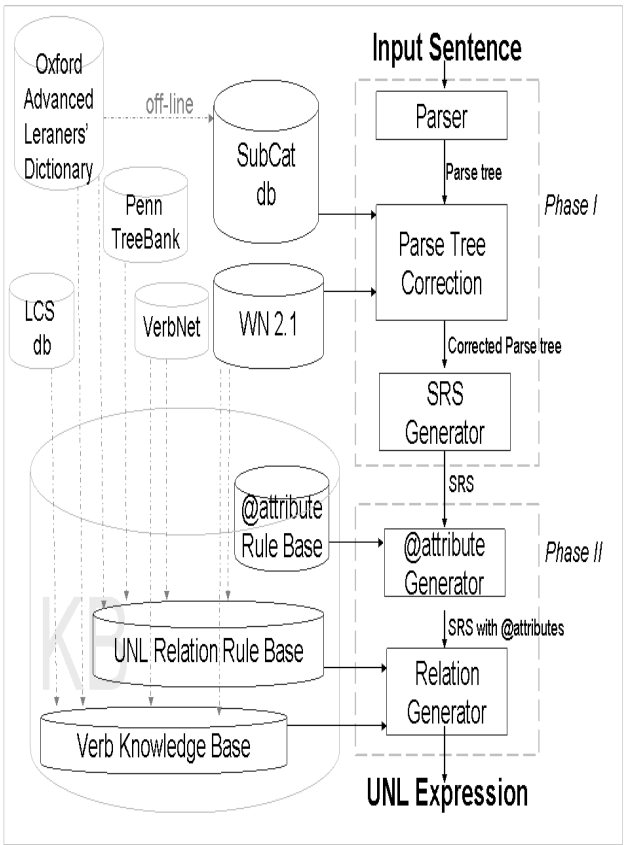


Figure. 3: The system architecture

## 5. Experimental Results

We show in figure 4 the accuracies of (i) SRS generation and (ii) UNL generation from SRS. On the

x-axis are the labels denoting corpora of various kind. For example, the first 10 3-member bar groups are for corpora from the OALD which are instances of controlled corpora in the sense of concentrating on particular language phenomena like Gerunds, Verbal Nouns, Present participles *etc*. The subsequent bars are for more open ended corpora, *e.g.*, from the *Times of India* news domain. We obtain an end-to-end sentence to UNL conversion accuracy of 54% (F1-score) which is reasonable considering the complexity of the UNL relation repository.

In one of our previous work (Mohanty et. al., 2005), we obtained an F1-score of about 0.6 for generating SRS from source English sentences.. This figure has been considerably improved upon to about 0.8 in our current system. This justifies the investment in lexical resources whose pay-offs are indeed considerable.

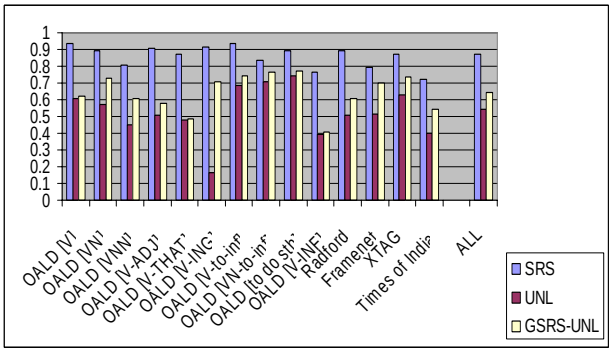


Figure. 4: F1-score for SRS, Sentence to UNL and gold-SRS to UNL generation

## 6. Conclusion and Future work

The aim of our research has been to establish a case for investment in lexical resources. Towards this, we have been able to demonstrate the utility of important lexical resources in the form of subcategorization databases, lexico-syntactico-semantic information for verbs and rules bases for setting up semantic relations and attributes. These resources no doubt are pain staking and need rare linguistic expertise. But once constructed, they prove their utility by way of being employed in solving a series of problems.

Our future work consists in increasing the quality and coverage of these resources.

## References


Chomsky, Noam. (1981). Lectures on Government and Binding. Foris, Dordrecht.

Dorr, Bonnie. (1992/1993). The use of lexical semantics in Interlingua Machine Translation, Machine Translation, 4/3.

Gruber, J.S. (1965) Studies in Lexical Relations. PhD Dissertation, MIT, Mass.

Hornby, A. S. (2001). Oxford Advanced Learners' Dictionary of Current English. OUP, 2001.

Jackendoff, Ray. 1990. Semantic Structures. The MIT Press, Cambridge.

LDC. (1995). Penn Treebank Release II. Linguistic Data Consortium.

Levin, Beth. (1993). English verb Classes and Alternation. The University of Chicago Press, Chicago.

Miller, George. (2005). WordNet 2.1. http://wordnet.princeton.edu/

Mohanty, Rajat., Anupama Dutta and Pushpak Bhattacharyya. (2005). Semantically Relatable Sets: Building Blocks for Knowledge Representation. Proceeding of 10th MT Summit, Phuket, Thailand.

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania.

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

UNDL Foundation. (2006). The Universal Networking Language (UNL) specifications (2006) http://www.undl.org

Woods, W. A. (1975). What's in a Link: Foundation for Semantic Networks, in Readings in Knowledge Representation, R.J. Brachman and H.J.Levesque (ed.), Morgan Kaufmann Publishers.

XTAG Research Group. (2001). XTAG Technical Report. University of Pennsylvania, Uppen.