ICON - 2008

6th International Conference on Natural Language Processing

Details Of The Selected Paper

Title	**Semantic Graph from English Sentences**
Topic	**Machine Translation**
Abstract	**In this paper we describe our progress towards building an interlingua based machine translation system, by capturing the semantics of the source language sentences in the form of Universal Networking Language (UNL) graphs from which the target language sentences can be produced. There are two stages to the UNL graph generation: first, the conceptual arguments of a situation are identified in the form of semantically relatable sequences (SRS) which are potential candidates for linking with semantic relations; next, the conceptual relations such as instrument, source, goal, reason or agent are recognized, irrespective of their different syntactic configurations. The system has been tested against gold standard UNL expressions collected from various sources like Oxford Advanced Learners' Dictionary, XTAG corpus and Framenet corpus. Results indicate the promise and effectiveness of our approach on the difficult task of interlingua generation from text.**
Authors	**Rajat Mohanty IIT Bombay Sandeep Limaye IIT Bombay M. Krishna IIT Bombay Pushpak Bhattacharyya IIT Bombay**
Contact	**pb@cse.iitb.ac.in**
Download	

Close

Semantic Graph from English Sentences

Abstract

In this paper we describe our progress towards building an interlingua based machine translation system, by capturing the semantics of the source language sentences in the form of *Universal Networking Language (UNL)* graphs from which the target language sentences can be produced. There are two stages to the UNL graph generation: first, the conceptual arguments of a situation are identified in the form of *semantically relatable sequences* (SRS) which are potential candidates for linking with semantic relations; next, the conceptual relations such as *instrument, source, goal, reason* or *agent* are recognized, irrespective of their different syntactic configurations. The system has been tested against gold standard UNL expressions collected from various sources like Oxford Advanced Learners' Dictionary, XTAG corpus and Framenet corpus. Results indicate the promise and effectiveness of our approach on the difficult task of interlingua generation from text.

Keywords: Universal Networking Language Expressions, Semantically Relatable Sets, Lexical Knowledge Bases, Syntactic and Semantic Constituents, Interlingua-based MT, Parse Trees, Lexical Knowledge Base.

1 Introduction

Unpacking Semantics is a key task in interlingua-based Machine Translation system. Our work is motivated by the *interlingua* called *Universal Networking Language (UNL)* (Uchida et. al., 2000). We aim at unpacking semantic information in terms of UNL graphs from English texts. We achieve the goal in two phases: (1) identifying the semantic arguments of a situation in terms of *Semantically Relatable Sequences (SRS)*, even when the arguments are expressed in different syntactic configurations; (2) assigning a UNL relation to each SRS in terms of *instrument, source, goal, reason, agent, etc.* Given an input sentence, the system breaks the constituents into one of the three basic semantically relatable sequence frames such as *<entity1 entity2>* or *<entity1 functor entity2>* or *<functor entity>*, where the entities can be single words or more complex sentence parts (such as embedded clauses). Ultimately, these sequences are labeled with either abstract semantic relations (like *agent (agt), object (obj), goal(gol), instrument (ins), source (src), etc.*), or are expressed in terms of attributes such as *@present, @past, @topic, @passive, @proximate, @interrogative, etc.* which are basically speek acts. In this system, we use a statistical parser (Charniak, 2004) and the extensive knowledge bases created offline taking help from various existing lexical resources such as, WordNet 2.1, LCS database (Dorr,), Oxford Advanced Learners' Dictionary (Hornby, 2001), VerbNet and Treebank (LDC, 1995).

Coming to related work, we stress that our work is ultimately an exercise in knowledge representation which has been extensively discussed in the classical treatises by Dorr (1992), Schank (1972) and Sowa (2000). Inerlingua representations have been studied in the machine translation literature (Hutchins and Somers, 1992). One of the early noteworthy interlingua based MT systems is Atlas-II (Uchida, 1989); the comparison of the interlingua approach to the more widespread transfer approach is done in Boitet (1988); the consequence of language divergence on interlingua has been recently studied in Dave *et. al.* (2002).

The roadmap of the paper is as follows: section 2 presents the UNL framework. Section 3 gives a rationale for using UNL. The notion of SRS and its relevance in the context of UNL is presented in Section 4. Section 5 describes the knowledge base forming the foundation of this work. Section 6 discusses the implementation. The experimental result is given in section 7. Section 8 concludes the paper.

2 Universal Networking Language: 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 (www.undl.org) 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. However, these tools are limited in their capability rely as they heavily on language expert's knowledge and intuitions. We describe here a robust and scalable approach based on syntactic analysis and exhaustive knowledge bases for UNL generation. The constituents of the UNL system are described now (UNDL, 2005).

2.1 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. 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.

2.2 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*.

2.3 UNL Relations

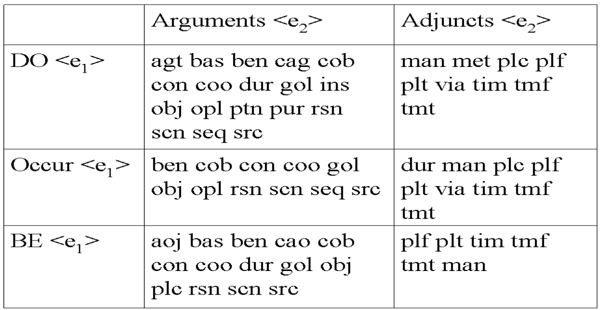
Binary relations of the UNL expressions represent directed binary relations between the concepts of a sentence. There are a total of 46 relation labels defined in the UNL specifications (UNDL, 2006). The syntax of Binary relations is as follows:

<Binary Relation>::= <Relation Label>[":"< Compound UW-ID>] "("<UW1>| ":" <Compound UW-ID1>","<UW2>| ":" <Compound UW-ID2>")"

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

2.4 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]

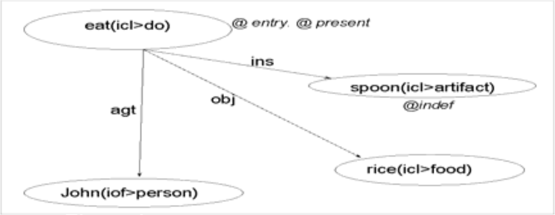


Figure 1: UNL graph of *John eats rice with a spoon*

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.

2.5 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 chunk "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. These UNL relations are depicted pictorially through a UNL graph as shown in Figure 2. Any compound concept can be represented using a scope and the scope technique allows us to capture deeply nested constructs in the language.

3 Why UNL?

In 1992, KANT (Nyberg *et. al.*, 1992)- the interlingua and the system with this name- was designed for large scale MT of technical documentation from English to a number of other languages. However, KANT is a sublanguage system, *i.e.*, it handles only a subset of English called *constrained technical English.*

UNITRAN- again the interlingua and the MT system with the same name- is too detailed a framework for any meaningful practical implementation (Dorr, 1992/93]). ULTRA (Farwel *et. al.*, 1991) - the American MT effort using interlingua- uses Prolog based grammar for the intermediate representation and is necessarily restricted in its scope for handling language phenomena.

UNL has been influenced by a number of linguistics-heavy interlingua based Japanese MT systems in the 1980s- notably the ATLAS-II system of Fujitsu (Uchida, 1989). However, the presence of a number of researchers from Indo-Iranian, Germanic and Baltic-Slavic language families in the committee for UNL specifications (Uchida *et. al.*, 1999) since 2000, has lent UNL a much more universal character compared to the interlingua used in ATLAS-II.

Comparing and contrasting UNL with primitive based interlingua like Conceptual Dependency (schank,, 1972) and Conceptual Structures (Sowa, 2000), we observe that like UNITRAN, they too are too detailed to admit practical implementations. If Conceptual Dependency, UNITRAN, Conceptual Structure are too fine-grained, the Esperanto like interlingua used in the Distributed Language Translation project conducted at the BSO company at in the Netherlands (Witkam, 1988, Schubert, 19888) is too coarse grained and fraught with ambiguity. Esperanto had the ambitious aim of being a universal language for people-to-people communication. UNL is a fine balance between the two extremes represented by UNITRAN and Esperanto.

We find that the UNL representation has the right level of expressive power and granularity. Additionally, we believe that for those working in a rich and diverse multilingual setting, *e.g.*, India, UNL provides the right representation for interlingual MT among Indian languages.

A comparison with the famed framenet project (Gildea and Jurafski 2002) is in order here. The Framenent project decided on hundreds of semantic roles which are more like frame elements rather than thematic roles (*i.e.*, roles relating nouns to verbs). The complex expressions are often assigned a single Framenet semantic role ignoring the crucial linguistic information involved in each and every thematic elements of that expression. For instance, a relative clause along with its antecedent is assigned a single semantic role. UNL on the other hand has 46 semantic relations which are mostly thematic roles assigned to each and every thematic element of an expression. In our understanding Framenet roles are suitable for information extraction tasks. A complex task like MT needs to capture and represent the relation between the verb and its arguments/adjuncts accurately.

UNL based semantic relation identification is thus a much more involved task than any of the existing ones we know.

4 Notion of Semantically Relatable Sequence (SRS)

In this section, we briefly look at the categorization of words and the possible association among them to identify the semantic arguments of a situation in terms of *Semantically Relatable Sequences* (Mohanty *et. al. 2005*), which, in turn, are used for UNL graph generation.

4.1 Semantically Relatable Sequences (SRS)

Sentence structures are usually divided into three functional domains: (i) a lexical domain around the verb, which establishes semantic relations between the main sentence elements; (ii) a grammatical domain around the auxiliary, which establishes grammatical relations such as agreement; (iii) a discourse domain around the complementizers and other elements, which links two clauses.

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.* That is, a sentence needs to be broken into sequences of at most three forms, which are referred to as SRS, as given in (5).

(5) a. (CW, CW)
b. (CW, FW, CW)
c. (CW, FW)

The notation FW stands for function words; CW stands either for a content word or for a clause. The concept of SRSs can be well understood by considering the example sentence in (6).

(6) The Professor spoke to the students in the lounge on Friday.

(7) a. (the, Professor)
b. (Professor, spoke)
c. (spoke, to, students)
d. (the, students)
e. (spoke, in, lounge)
f. (the, lounge)
g. (spoke, on, Friday)

Our objective is to use a syntactic form as the starting point for generating a semantic representation. Hence, we show the intuition behind deriving these SRSs from a syntactic analysis output, *i.e.*, the parse tree. In the parse tree, we treat tags like NP, and VP as tags indicating the presence of content words and tags like PP (prepositional phrase), IN (preposition) and DT (determiner) as denoting function words. Consider the partial parse tree for the sentence (10) given in Figure 2, in which the (C) and (F) tags denote content words and function words, respectively, and the subscripts indicating the head words of the subtrees dominated by the nodes. It is observed that most SRSs are constituted of head words of sibling nodes.

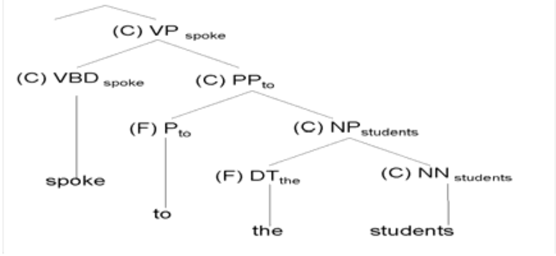


Figure 2: Partial Parse Tree for *spoke to the student*

Once a sentence is broken up into SRS, no structural ambiguity is expected to be left for resolution. Subsequently, each SRS safely either leads to the generation of a semantic relation or is translated into the UNL attribute labels indicating the subjectivity of the sentence, depending upon the kind of elements present in a particular sequence, as illustrated in (8).

(8) John has cut the cake with a knife.

[SRS:8]
(John, cut.@entry)------------(CW, CW)
(has, cut.@entry)--------------(FW, CW)
(cut.@entry, cake)------------ (CW, CW)
(the, cake)---------------------- (FW, CW)
(cut.@entry,with, knife)------(CW, FW, CW)
(a, knife)-------------------------(FW, CW)
[\SRS]

[UNL:8]
agt(cut.@entry.@present.@complete, John)
obj(cut.@entry.@present.@complete, cake.@def)
ins(cut.@entry.@present.@complete, knife.@indef)
[\UNL]

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
Table 2														

4.2 SRS and Different Language Phenomena

The generation of SRSs often needs solving different kinds of problematic language phenomena that include the attachment issues (*such as*, PP-attachment and clausal-attachment), detection of empty pronominals in non-finite clauses (*i.e.*, to-infinitival and gerundial clauses), detection of movement traces (as found in interrogatives, topicalization, PP-stranding, relative clauses), and other specific issues pertaining to copular constructs, small clauses, partitive constructs, among others.

5 Knowledge Base (KB)

We have built an exhaustive knowledge base for UNL generation. Basically it consists of Subcategorization KnowledgeBase, Verb KnowledgeBase, UNL Relation RuleBase, and UNL @attribute RuleBase. On the whole, it provides linguistic knowledge of concepts, argument frames, subcategorization details, semantic features of lexical elements, tense-aspect details along with some pragmatic information.

5.1 Lexical Subcategorization Knowledge

Every content word, whether it is noun, verb, or adjective, has certain elements that it is said to subcategorize (Chomsky, 1981). Lexical items can subcategorize prepositions and clauses of different kinds. This subcategorization information is obtained from the Oxford Advanced Learner's Dictionary (OALD) (Hornby, 2001) by collecting the details manually.

5.2 UNL Relation RuleBase

The Relation Rule Base is one major component of the knowledgebase in the system. For each rule, a specific relation generation template is used. The existing linguistic resources like VerbNet (Schuler, 2005), WordNet 2.1 (Miller, 2005) , OALD (Hornby, 2001) , Treebank (LDC, 1995) are used off-line in developing the RuleBase. The number of rules is about 700 at present, and is getting enriched day-by-day with hand-crafted rules acquiring from the above resources. The rule template is depicted in Table 2.

4.2.1 Syntactic Features

The field Syntactic Feature in the rule template consists of two subfields, such as syntactic category (SynCat) and Parts-Of-Speech (POS). The *SynCat* field is defined to be one of the lexical categories, such as N, V, J, R and P referring to nouns, verbs, adjectives, adverbs, and prepositions, respectively. The *SynCat* field is mapped to the POS field considering the parser generated POS tags.

4.2.2 Semantic Features

The field Semantic Feature in the rule template consists of two subfields, such as semantic category (*SemCat*) and the actual lexical item (*Lex*). The *Lex* field is filled only when it is very specific as in case of FW or when the *SemCat* field is not yet defined. The *SemCat* field is defined for verbs, nouns, and adverbs, so far.

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

There are about 139 tables storing approximately 3900 verbs. The rules containing one of these 139 classes are developed off-line from VerbNet data. The relevance of such a rule in UNL relation generation for the SRS *(cut, with, knife)* is illustrated below in terms of (CW1,FW,CW2) and the corresponding rule:

CW1	FW	CW2	Rel
V _ v139 _	P IN _ with	N _ _ _	Ins

Table 3: Rule for SRS having a Verb class

Nouns: The *SemCat* field for the nouns carries the semantic grouping of nouns on the basis of the WordNet 2.1 (Miller, 2005) noun classification. The semantic features like TIME, PLACE, ANIMATE, INSTRUMENT, LEGAL DOCUMENT, etc. (which are relevant in the context of UNL generation) are detected using the hypernymy hierarchy of the words in the WordNet. The rules specifying the *SemCat* for nouns are developed off-line from WordNet.

Adverbs: The The *SemCat* field for the adverbs carries the semantic grouping of adverbs on the basis of the classification done in the Penn TreeBank Release II (LDC, 1995). The lexical items having the tags like ADV-MNR, ADV-TMP and ADV-LOC are acquired from the Treebank, and are encoded in the RuleBase. The following table illustrates a few rules for the SRSs containing adverbs *(playing, there), (coming, early), (playing, there)*:

CW1	FW	CW2	Rel
V _ _ _	----	R _ MNR _	man
V _ _ _	----	R _ TMP _	tim
V _ _ _	----	R _ LOC _	plc

Table 4: Rules for SRS having an Adverb

5.3 Verb Knowledge Base (VKB)

The Verb Knowledge Base(VKB) contains the lexical, syntactic and semantic information associated with a verbs. It is created offline using various linguistic resources like WordNet 2.1 (Miller, 2005) and Lexical Conceptual Structure database (Dorr, 1992) and OALD (Hornby, 2001). Each entry in the knowledgebase is illustrated in Table 5:

V	icl>UW	BE|DO| OCCUR	<AF>	REL: ARG-TYPE
give	icl>supply	do	agt>thing, obj>thing, gol>thing	gol:to
give	icl>transfer	do	agt>thing, obj>thing, gol>thing, src>thing	gol:to, src:from

Table 5: Verb KnowledgeBase Structure

The initial phase of the creation of the Verb KB was done taking LSC database (Dorr, 1992) as the prime source covering around 3000 verbs. Subsequently, in the later phase more verb entries are being studied, looking at the WordNet, the OALD, and other relevant resources, and in turn, they are added manually to the VKB. The current coverage of unique verbs is 6,298 and the number of corresponding argument frames is 46,134. The number is increasing day-by-day with hand-crafted verb entries.

5.4 UNL @attribute RuleBase

The function words (FWs) in the sentences get associated with the CWs in the form of SRSs of the type {FW, CW}. These SRSs are processed to generate UNL @attributes for the concerned CWs. For example, the @passive attribute is generated from sequences of the form (<be-aux>, VBN) type SRS as found in the sentence like *This letter must have been written by her*. There are different combinations of modals, auxiliaries and verb-forms (such as, VBD, VBN, VBZ, etc.) specified in the rule base to contribute to the generation of different UNL attributes for verbs. Similarly, there are rules to generate attributes for nouns and adjectives.

String of FWs	CW	@attribute list
has_been	VBG	.@present. @complete. @progress
has_been	VBN	.@present. @complete. @passive

Table 6: @attribute RuleBase Template

6 Implementation

The design and implementation of the UNL generation system is done with a focus on flexibility and extensibility. The most vital and valuable component of this system is its knowledgebase, which is expected to be improved as the linguistic insights and perceptions change over time. Keeping this in mind, the database tables have been designed to be as independent of each other and the code as well. The database tables are easily modifiable and extensible, leaving room for improvement.

6.1 System Architecture

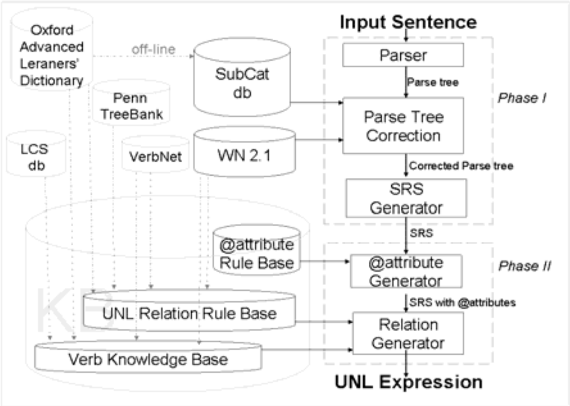


Figure 3: The System Architecture

6.2 Overall Strategy

SRS Generation

Step 1: Get the parsed output from charniak.
Step 2: Build a tree data structure.
Step 3: Identify heads.
Step 4: Generate SRSs of the patterns
(FW,CW), (CW,CW), (CW,FW,CW)

SRS-to-UNL Generation

Step 1: Accept the SRS intput.
Step 2: Generate attributes using (CW,FW) pairs or tags of CW.
Step 3: Split the SRSs into VerbBased, Non-Verb Based triplets.
Step 4: Generate relations for non-verb based SRSs using Rule base.
Step 5: Other than the basic 8 synatctic frames solve all the other arguments of each verb as adjuncts.
Step 6: Solve the basic verb structures (For each of the structure the recursive strategy is used.

6.3 Recursive Strategy

Theoretically, a verb or a noun can legitimately take a fixed number of arguments (possibly maximum three) but innumerable adjuncts. However, we studied all the possible syntactic frames in the Treebank (LDC, 1995), in which we found that there exists maximum seven post-verbal argument-adjunct positions for verbs. Out of about 3000 different syntactic frames (for verbs), we devised the following 8 steps as the recursive strategy for UNL generation.

Step 1 [N_0 -V]

```
a.  If V has @passive, then assign  obj(V,$N_0$
    )
b.  Else
    determine the verb group info,
      if $V_{unErgBe /vEcm}$ then aoj(V, $N_0$)
      else if $V_{unErgDo}$ then agt(V, $N_0$)
      else if $V_{erg}$ then obj(V, $N_0$)
      else if $V_{@animate}$ then agt(V, $N_0$)
      default: obj(V, $N_0$)
```

Step 2 [N_0 –V-AP]

```
a.  If SRS is (C,F,C) and the V is {is, am,
    are, was, were, be, been, being},  assign
    aoj(AP,$N_0$)
b.  default: aoj($V_{BE}$,$N_0$), gol($V_{BE}$,AP)
```

Step 3 [N_0 –V-PP]

```
a.  Resolve PP using RuleBase
b.  If generated relation is found in <VKB>,
      take the argument structure from <VKB>
c.  Else follow Step 1
```

Step 4 [N_0 –V-N_1]

```
a.  If $N_1$ has [PLACE]/[TIME] ,
  (i)   resolve $N_1$  with
        plc|opl|tim|dur
  (ii)  look up <VKB>,
        If the generated relation is found in
    <VKB>, resolve $N_0$
        Else follow Step 1
b.  Else look up <VKB>
  (i)   If only one frame with 2 roles is
        found in <VKB>, resolve $N_0$ and $N_1$ ac-
        cordingly.
  (ii)    Else (default)
        agt(V , $N_0$), obj(V , $N_1$)
```

Step 5 [N_0 –V-N_1-PP]

```
a. Resolve PP using RuleBase
b. If generated relation is found in <VKB>,
    take the argument structure from <VKB>,
   else follow Step 1
```

Step 6 [N_0-V-N_1-N_2]

```
a. If $N_2$ has [PLACE]/[TIME],
  (i) resolve $N_2$  with plc/tim/dur
  (ii)look up <VKB>,
      if plc/plf/tim/dur is found in
        <VKB>, resolve $N_0$ and $N_1$
      else follow Step 4
b.Else if single frame with 3 roles
    is found in <VKB> , resolve $N_0$,$N_1$,and $N_2$
```

```
c.  Else (default)
    agt(V,$N_0$), gol(V,$N_1$), obj(V,$N_1$)
Step 7  [$N_0$-V- S/SBAR]
a. use RuleBase to resolve the
    S/SBAR
b. if the generated relation is
    found in the <VKB>, Resolve $N_0$
    Else follow Step 1
Step 8  [$N_0$-V-$N_1$ -S/SBAR]
a. use RuleBase to resolve the
    S/SBAR
b. if the generated relation is
    found in the <VKB>, Resolve $N_0$,$N_1$
Else follow Step 4
```

7 Experimental Results

7.1 Creation of Test data

We created the testbed taking example sentences from various authentic sources like XTAG Technical Report (XTAG, 2001), OALD (Hornby, 2000), FrameNet Book (Ruppenhofer *et. al.*, 2005), and Transformation Grammar (Radford, 1998), in which a wide range of language phenomena are presented. Out of all the example sentences available in these resources, 504 sentences are randomly picked up for the current evaluation, for which *gold standard* UNL have been created with manual effort.

7.2 Experiments and Top Level Statistics

The UNL expressions generated by our system were compared with the gold standard UNL expressions. We are inspired by Information Retrieval in assigning recall and precision values to these comparisons, where recall, precision and the F1 score are defined as given below.

$$Score_{UNL}(unl_{Generated}, unl_{Gold}) = \frac{2 * precision * recall}{precision + recall}$$

$$precision = \frac{\sum_{\forall unle \in unl_{Generated}} Score_{UNLE}(unle)}{count(unle \in unl_{Generated})}$$

$$recall = \frac{\sum_{\forall unle \in unl_{Generated}} Score_{UNLE}(unle)}{count(unle \in unl_{Gold})}$$

$$\begin{aligned} Score_{UNLE}(unle) = Average(&Score_{Relation}, \\ &Score_{UW}(uw1_{unle}), Score_{UW}(uw2_{unle})) \end{aligned}$$

$$\begin{aligned} Score_{Relation} &= 1 : \text{ if generated relation name is correct} \\ &= 0 : \text{ otherwise} \end{aligned}$$

$$Score_{UW}(uw) = Average(Score_{Word}, Score_{attributes})$$

$$\begin{aligned} Score_{word} &= 1 : \text{ if generated lexical word is correct} \\ &= 0 : \text{ otherwise} \end{aligned}$$

$$Score_{attributes} = F1Score(Attributes_{Generated}, Attributes_{Gold})$$

7.3 Example of Applying Evaluation Formula

```
Sentence: He worded the statement carefully.
[unlGenerated:76]
agt(word.@entry, he)
obj(word.@entry, statement.@def)
man(word.@entry, carefully)
[\unl]
;He worded the statement carefully.
```

```
[unlGold:76]
agt(word.@entry.@past, he)
obj(word.@entry.@past, statement.@def)
man(word.@entry.@past, carefully)
[\unl]
Score_unl = 2(precision *
        recall)/(precision+recall) =
precision = sum(0.945,0.945,0.945)/ 3
        = 0.945
recall  = sum(0.945,0.945,0.945)/ 3 = 0.945

Score_unle(agt(word.@entry, he))
        = average(1, 0.835, 1) = 0.945
Score_unle(obj(word.@entry, statement.@def))
        = 0.945
Score_unle(man(word.@entry, carefully))
        = 0.945
Score_relation = 1 for all relations of
        unle(s) here
Score_uw(word.@entry)
                = average(1,0.67)=0.835
Score_word = 1 for all words of unle(s) here
Score_attributes = 2 (1*0.5)/(1+0.5)  = 0.67
```

	Precision	Recall	F1 score
XTAG	0.632	0.618	0.624
FrameNet	0.685	0.663	0.672
TG	0.725	0.718	0.720
OALD	0.523	0.497	0.508
Overall	0.622	0.604	0.611

Table 7: Statistics for NL text to UNL generation

F1 score for SRS-to-UNL = 0.788

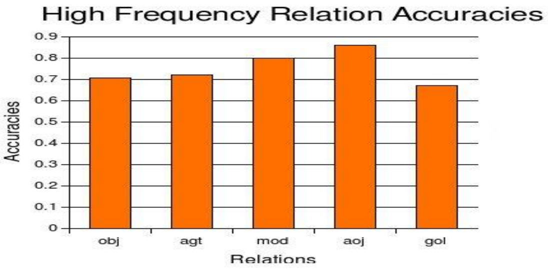


Figure 4

8 Conclusion and Future work

We have reported here a robust and scalable method for semantic representation generation with reasonable high accuracy (61%). The work reported is part of an MT effort involving interlingua. Some of the important stuffs are not reported here due to lack of space. The investigation also underlines the importance of designing rich and high-quality knowledgebase. Our future work mainly concentrates on the enrichment of knowledgebase as well as the possibility of using a high accuracy parser as a starting point (*e.g.*, LFG Grammar and XLE parser).

References

Boitet Christian. 1988. Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation. In Klaus Schubert and Toon Witkam (eds.), Recent Developments in Machine Translation. Dan Maxwell, Foris, Dordrecht.

Charniak, Eugene, Don Blaheta, Niyu Ge, Keith Hall, John Hale and Mark Johnson. *WSJ Corpus Release 1*. LDC.

Dorr Bonnie. 1992.. The use of lexical semantics in Interlingua Machine Translation, *Machine Translation*, 7.

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

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

Dave Shachi, Jignashu Parikh and Pushpak Bhattacharyya. 2002. Interlingua Based English Hindi Machine Translation and Language Divergence, Journal of Machine Translation (JMT), 17.

Gildea D. and Jurafsky D..2002. Automatic Labeling of Semantic Roles. Computational Linguistics, Vol. 28, No. 3.

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

Farwel D. and Y. Wilks. 1991. ULTRA, a Multilingual Machine Translator, MT Summit III , Washington, DC, USA.

Karin Kipper Schuler. 2005. VerbNet: A broad-

coverage, comprehensive verb lexicon. University of Pennsylvania.

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.

Nyberg E. and Mitamura T. 1992. The KANT system: Fast, accurate, high-quality translation in practical domains. In Coling-92.

Radford Andrew. 1998. Transformation Grammar. CUP.

Ruppenhofer Joseph, Michael Ellsworth, Miriam R. L.,Petruck, Christopher R. Johnson. 2005. *FrameNet: Theory and Practice*.

Schank Roger C. 1972. Conceptual Dependency: A Theory of Natural Language Understanding. Cognitive Psychology, 3.

Schubert, K. (1988). The Architecture of DLT-interlingual or double-dialect, in New Directions in Machine Translation, Floris Publications, Holland.

Sowa John F. 2000. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks/Cole Publishing Co., Pacific Grove, CA.

Uchida Hiroshi. 1989. ATLAS-II: A machine translation system using conceptual structure as an interlingua. In Proceedings of the Second Machine Translation Summit, Tokyo.

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 http://www.undl.org

XTAG Research Group. 2001. XTAG Technical Report. University of Pennsylvania, Uppen. 29. UNDL Foundation. (2006). The Universal Networking Language (UNL) specifications (2006) http://www.undl.org

Witkam, T. 1988. DLT- an Industrial R & D Project for Multilingual Machine Translation, COLING , Budapest.