Semantic Graph from English Sentences

Rajat Mohanty **Sandeep Limaye**
M. Krishna Prasad **Pushpak Bhattacharyya**
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai-400076, INDIA
Email: {rkm,sandeep,mkrishna,pb}@cse.iitb.ac.in

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.

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 grammatical attribute labels such as *@present, @past, @topic, @passive, @proximate, @interrogative, etc.* In this system, we use a statistical parser (Charniak, 2004) and the extensive knowledge bases created off-line taking help from various existing lexical resources such as, WordNet 2.1, LCS database (Dorr,), Oxford Advanced Learners' Dictionary (Hornby, 2001), VerbNet (Schuler, 2005) 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). Interlingua 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

Proceedings of ICON-2008: 6th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008

relevance in the context of UNL is introduced in Section 4. Section 5 introduces 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>]

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. 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]

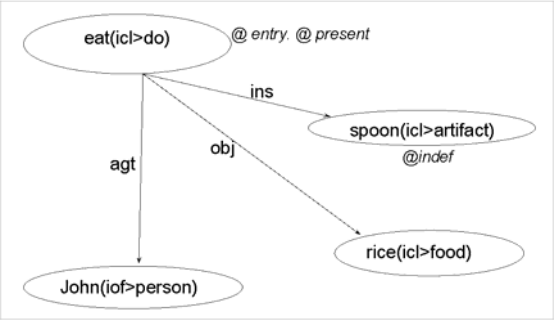

Figure 1: UNL graph of *John eats rice with a spoon*

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.

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 FrameNet 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 (such as content words (CW) and func-

tion words (FW)) 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. Our objective is to use a syntactic form as the starting point for generating a semantic representation. 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.

5 Knowledge Base (KB)

We have built an exhaustive knowledge base for UNL generation, described in (Mohanty and Bhattacharyya, 2008). The knowledgebase 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.

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 Overall Strategy

SRS Generation

Step 1: Get the parsed output from charniak parser.
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 input.
Step 2: Generate attributes using (CW, FW) pairs or tags of CW.
Step 3: Split the SRSs into Verb Based, Non-Verb Based triplets.
Step 4: Generate relations for non-verb based SRSs using the Rule base.
Step 5: Other than the basic 8 syntactic 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.

6.2 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 accordingly.
(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 N0 and N1
      else follow Step 4
b. Else if single frame with 3 roles
    is found in <VKB> , resolve N0,N1,and N2
c.  Else (default)
    agt(V,N0), gol(V,N1), obj(V,N1)
```

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 N0
    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 N0,N1
Else follow Step 4
```

6.3 System Architecture

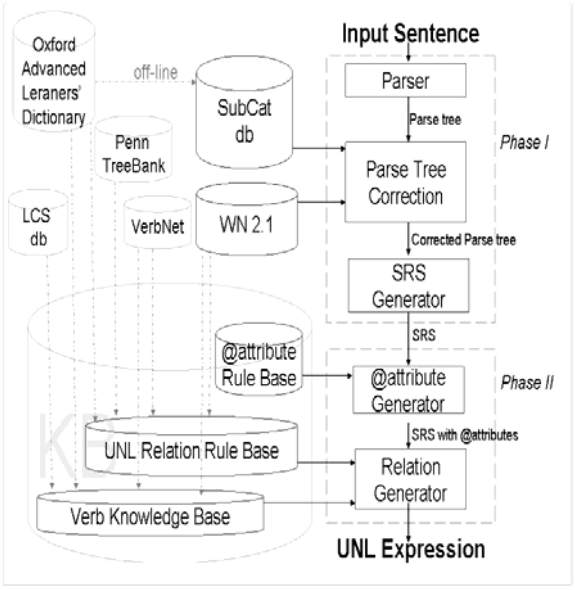

Figure 2. The System Architecture

7 Experimental Results

7.1 Creation of Test data

We created the test bed taking example sentences from various authentic sources like XTAG Technical Report (XTAG, 2001), OALD (Hornby, 2000), FrameNet II (Ruppenhofer *et. al.*, 2006), 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 Evaluation Formula

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.

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

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

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

$$Score_{UNLE}(unle) = Average(Score_{Relation}, Score_{UW}(uw1_{unle}), Score_{UW}(uw2_{unle}))$$

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

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

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

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

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
```

7.4 Top Level Statistics

	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 2. Statistics for NL text to UNL generation

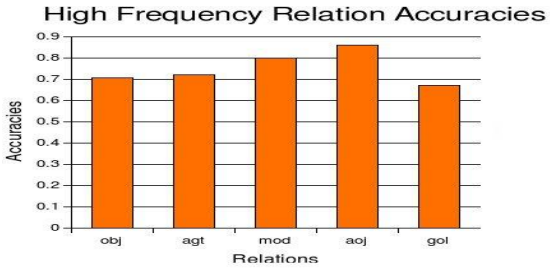

Figure 3. Accuracy for high frequency relations

8 Conclusion and Future work

We have reported here a robust and scalable method for semantic representation generation with reasonable high accuracy (61%). The F1 score for GoldSRS-to-UNL is as high as 78%. 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

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

Andrew Radford. 1998. Transformation Grammar. CUP.

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

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

Christian Boitet. 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.

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

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

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

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

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

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

George Miller. 2005. Wordnet 2.1. http://wordnet.princeton.edu/

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

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

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

Joseph Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Jan Scheffczyk. 2006. FrameNet II: Extended Theory and Practice. http://framenet.icsi.berkeley.edu/book/book.html

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

Karin Kipper Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania.

LDC, 1995. Penn Treebank Release II. Linguistic Data Consortium.

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

Rajat Mohanty and Pushpak Bhattacharyya. 2008. Lexical Resources for Semantic Extraction. Proceedings of The 6th Language Resources and Evaluation Conferences (LREC 2008), Morocco.

Rajat Mohanty, Anupama Dutta and Pushpak Bhattacharyya. 2005. Semantically Relatable Sets: Building Blocks for Knowledge Representation. Proceedings of the MT Summit X, Phuket, Thailand.

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

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

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

XTAG Research Group. 2001. XTAG Technical Report. University of Pennsylvania, Uppen. 29.