Predicate Preserving Parsing

Jignashu Parikh Jagadish Khot Shachi Dave Pushpak Bhattacharyya[1]
Department of Computer Science and Engineering,
Indian Institute of Technology, Bombay.

Abstract

We present a unique approach to knowledge extraction from texts by a method of natural language analysis which *preserves the predicate till the end*. The system thus named *Predicate Preserving Parser* (PPP) performs morphological, syntactic and semantic analysis synchronously. This approach helps in highly accurate analysis of sentences. The analysis produces a semantic net like structure expressed by means of Universal Networking Language (UNL)- a recently proposed Interlingua. The working of the PPP is demonstrated through the analysis of English and Hindi sentences. Varied and complex phenomena of both the language have been tackled. Use of lexical resources like the WordNet facilitates the handling of language phenomena.

Keywords: Knowledge Extraction, Universal Networking Language, Natural Language Parsing, Predicate Preservation, Semantic Net.

1 Introduction

We describe a system to automate the generation of semantic net like expressions from text documents. The system reified as *Predicate Preserving Parser* (PPP) performs morphological, syntactic and semantic analysis synchronously. The objective is to establish appropriate relations between the syntactic units of a sentence. PPP uses the word and world knowledge, *i.e.*, syntactic and semantic attributes of words. The output of the system is a set of *Universal Networking Language Expressions,* which are binary relations among disambiguated words along with speech act attributes attached to these disambiguated words.

The Universal Networking Language (UNL 1998) has been defined as a digital meta language for describing, summarizing, refining, storing and disseminating information in a machine independent and human language neutral form. UNL represents information in a document sentence by sentence. Each sentence is converted to a directed hyper graph having concepts as nodes and relations as arcs. Knowledge within a document is expressed in three dimensions:

a. Word Knowledge is expressed by **Universal Words (UWs)** which are language independent. These UWs are tagged using restrictions describing the sense of the word in the current context. For example, *drink(icl>liquor)* denotes the noun sense of *drink* restricting the sense to a type of *liquor*. Here, *icl* stands for inclusion and forms an *is-a* relationship like in semantic nets (woods 1985).

b. Conceptual Knowledge is captured by relating UWs through a set of **UNL relations** (UNL 1998). For example, *Humans affect the environment* is described in the UNL as,

   agt(affect(icl>do).@present.@entry:01, human(icl>animal).@pl:I3)
   obj(affect(icl>do).@present.@entry:01, environment(icl>abstract thing).@pl:I3)

   agt means the *agent* and **obj** the *object. affect(icl>do), human(icl>animal)* and *environment(icl>abstract thing)* are the UWs denoting concepts.

c. Speaker's view, aspect, time of event, *etc.* are captured by **UNL attributes**. For instance, in the above example, the attribute *@entry* denotes the main predicate of the sentence, *@present* the *present tense* and *@pl* the *plural number.*

The above discussion can be summarized using the example below:

John, who is the chairman of the company, has arranged a meeting at his residence.

[1] Author for Correspondence, pb@cse.iitb.ac.in

UNL for the sentence is,

```
;======================== UNL =======================
;John who is the chairman of the company has arranged a meeting at his residence.
[S]
mod(chairman(icl>post):01.@present.@def,company(icl>institution):02.@def)
aoj(chairman(icl>post):01.@present.@def,      John(icl>person):00)
agt(arrange(icl>do):03.@entry.@present.@complete.@pred,John(icl>person):00)
pos(residence(icl>shelter):04, John(icl>person):00)
obj(arrange(icl>do):03.@entry.@present.@complete.@pred,meeting(icl>conference):05.@indef)
plc(arrange(icl>do):03.@entry.@present.@complete.@pred,residence(icl>shelter):04)
[/S]
;===================================================
```

The UNL graph for the sentence is given in figure 1.

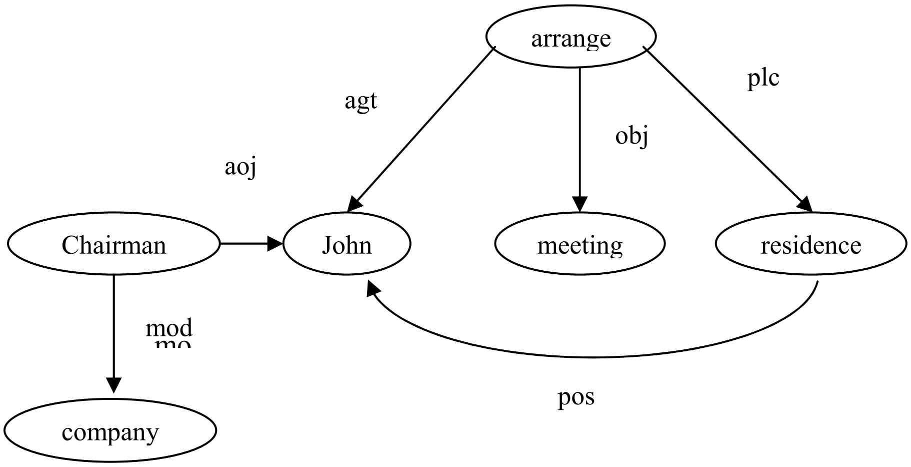

Figure 1: UNL graph

In the figure above, *agt* denotes the *agent* relation, *obj* the *object* relation, *plc* the *place* relation, *pos* is the *possessor* relation, *mod* is the *modifier* relation and *aoj* is the *attribute-of-the-object* (used to express constructs like *A is B)* relation. The detailed specification of the Universal Networking Language can be found at *http://www.unl.ias.unu.edu/unlsys.*

In the next section, we introduce the machine for generating the UNL. The Predicate Preserving Parser (PPP) is discussed in section 3. Section 4 lists the various English language phenomena handled. Section 5 is on Hindi analysis. The evaluation of the system is given in section 6. Section 7 gives the related work and the comparison. Concluding remarks follow in Section 8. UNL expressions for some example sentences are given in the appendix.

2 The UNL Generating Machine: EnCo

The EnConverter (henceforth called *EnCo*) is an analyser tool provided by the UNL Project, Institute for Advanced Studies, United Nations University, Tokyo. It analyses texts sentence by sentence using a knowledge rich **lexicon** and interpreting the **analysis rules**. Structured as a multi-headed Turing Machine (figure 2 below), it moves back and forth over the *Node-list* which contains words of the input sentence. The heads of the EnCo are called *windows*. There are two kinds of windows, *viz., Analysis Windows (AW)* and *Condition Windows (CW)* CWs circumscribe the AWs and are used for checking the conditions on the nodes on both sides of the Analysis Windows.

EnCo is driven by the *analysis rules* which have the following syntax (EnCo 2000):

<TYPE> (<PRE>)… {<LNODE>} {<RNODE>} (<SUF$_1$>) (<SUF$_2$>) (<SUF$_3$>)… P<PRI>;

Where,

<LNODE>:="{“ [<COND1>] ":" [<ACTION1>] ":" [<RELATION1>] "}"

<LNODE>:="{“ [<COND2>] ":" [<ACTION2>] ":" [<RELATION2>] "}"

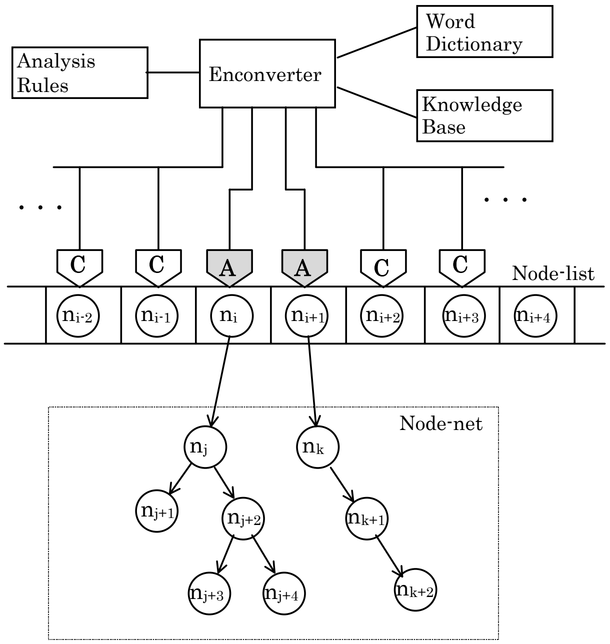

Figure 2: Structure of EnCo
“A” indicates an Analysis Window, “C” indicates a Condition Window,
and “n_n” indicates an Analysis Node

LNODE stands for the node under the Left Analysis Window and RNODE for that under the Right Analysis Window. The meaning of the above is:

> **Under the Left Analysis Window there is a node that satisfies <COND1> attributes, and under the Right Analysis Window there is a node that satisfies <COND2> attributes. When there are nodes to the left of the LAW, between the LAW and the RAW and to the right of the RAW that fulfil the conditions in <PRE>, <MID> and <SUF> respectively, the Lexical attributes in the nodes under the Analysis Windows are rewritten according to <ACTION1> and <ACTION2> respectively. Operations are done on the Node-list depending on the type of the rule shown in the field <TYPE>.**

<RELATION> fields are used to produce UNL relations between the nodes under the Analysis Windows. <PRI> indicates the priority value of the rules, which is in the range of 0-255. **The central task in creating a natural language analyser using the EnCo is to build a rich lexicon and a comprehensive set of analysis rules.**

EnCo uses a semantically rich lexicon of which some example entries are:

[bird] {}"bird (icl>animal>animate thing)" (N, ANI, SG, CONCRETE);
[beautiful]{} "beautiful(icl>state)"(ADJ);
[try] {} "try(icl>do)" (V, PRES, SIMPL)s;

Here *N* stands for noun, *ANI* for animate object, *SG* for singular, *concrete* for concrete object, *ADJ* for adjective, *V* for verb, *PRES* and *SIMPLE* for simple present. There are about 75 morphemic, syntactic and semantic attributes. The last mentioned and the restrictions on the UWs are attempted to be automatically extracted from the WordNet ontology (Verma and Bhattacharyya 2002).

3 The System

The idea of *predicate preservation* plays a major role at every step of analysis *where the attempt is always to locate the main predicate of the sentence***.** A node is deleted from the Node-list at every combination or modification operation if the node under the other Analysis Window is a better candidate for the predicate of the sentence or the clause. For example, if there is a proper verb under the LAW (by proper verb, we mean a verb that is not an auxiliary or modal verb) and a noun- which is an object- under the RAW then we delete the noun and preserve the verb. Similarly, if there is an auxiliary verb like *is* under the LAW and a noun under the RAW then we preserve the noun.

3.1.1 Overall Strategy

As the EnCo scans the input sentence left to right, two actions take place every step, (i) *morphological analysis (inflexional)* and (ii) *decision making*. The latter involves deciding, according to the attributes of the nodes under the two Analysis Windows, whether (i) the nodes are to be combined into a single headword or (ii) a relation is to be set up between them and/or (iii) an UNL attribute is to be generated. While combining or modifying the two nodes, one of the nodes is deleted from the node-list.

Multiple rules may become eligible for firing in a situation, calling for assignment of priorities for the rules as in expert systems. The strategy for prioritising the rules is briefly as follows:

i. Morphological analysis rules have the highest priority. Obviously, unless we have the morphed word we cannot decide the part of speech of the word and its relation with the adjacent words.

ii. Rules for dealing with specific constructs are given higher priority than those for general sentence structures. For instance, rules for clausal and passive sentences are given higher priority, so that while analysing clausal or passive sentences a general rule- eligible to be applied- does not fire.

iii. Right shift rules which facilitate right movement when there is nothing else to do are given the lowest priority. For example, when the LAW is on SHEAD (sentence start marker) and the RAW is on the subject (N), no rule other than the right shift is applicable. This rule, which is very useful is

```
R{SHEAD:::}{N:::}P1
```

iv. Composition rules are usually given less priority than modification rules. The former ultimately resolve relations while the latter change the properties of the nodes under the AWs.

3.1.2 Illustration of working

We illustrate the steps of the analysis process using a simple assertive sentence (only the major steps). Assertive simple sentences have only one main clause. The analysis strategy is explained below with an example where the node list is shown within "<<" and ">>" and the Analysis Windows are shown within "[" and "]". The nodes delimited by "/" are those visited and analysed by the machine. At every step, of the two nodes under the analysis windows the one with less importance is deleted. By *less important,* we mean that the particular node is not so much a candidate for being the main predicate of the sentence as the other one.

The sentence under consideration is *A report of John's genius reached the king's ears.*

```
<<[A ][report ]of John's genius reached the king's ears>>
```

The article and the noun are combined and the attribute *@indef* is added to the noun. The article *a* is relatively less important than the noun *report* and is thus deleted.

```
<<[report ][of] John's genius reached the king's ears>>
```

EnCo right shifts to put the preposition *of* together with the succeeding noun. This is based on the observation that a Noun Phrase is succeeding a preposition.

```
<</report /[of ][John's] genius reached the king's ears>>
```

John's being a possessive form we shift right to find its complement (*genius*).

```
<</report //of / [John's] [genius] reached the king's ears>>
```

These two nouns are resolved into the relation **mod** (meaning *modifier*) and the first noun is deleted as the sentence speaks about the *genius* of *John* and not about *John* himself.

```
<</report /[of][genius] reached the king's ears>>
```

The preposition *of* is then combined with the noun and a dynamic attribute OFRES (*of* resolved) is added to the node of *genius*.

```
<<[report][of genius ] reached  the king's ears>>
```

Using OFRES, the semantic relation between the two nouns are resolved into **mod** and the second noun is deleted.

```
<<[report ][reached] the king's ears>>
```

Shift right again and solve *the king's ears* as above. But now the relation is **pof** (part of) and not *pos* or *mod*. This distinction is achieved by the semantic attribute POF-ANI (part-of animate being) for *ear* present in the dictionary. The morphological analysis of the node *reached* attaches to it the attributes *@past* and *@pred*.

```
<</report /[reached ][ears]>>
```

The verb *reached* and the noun *ears* are resolved into **obj** and the noun is deleted.

```
<<[report ][reached ]>>
```

The noun *report* and the verb *reached* are resolved to **agt** (used for *agent*) and the noun is deleted again.

```
<<[reached ][>>]
```

A right shift at this point brings the Sentence Tail (STAIL) under the Left Analysis Window and thus signals the end of the analysis process. This right shift rule also attaches the attribute *@entry* to the last word left in the Node-list and thus the predicate *reached* is preserved till the end. The UNL produced is,

```
;=============================== UNL ============================
;A report of John's genius reached the king's ears.
[S]
mod(genius(icl>capacity):0I, John(icl>person):0C)
mod(report(icl>document):02.@indef,            genius(icl>capacity):0I)
pof(ear(pof>body):14.@pl,    king(icl>emperor):0X.@def)
agt(reach(icl>event):0P.@entry.@pred.@past,            report(icl>document):02.@indef)
plc(reach(icl>event):0P.@entry.@pred.@past,            ear(pof>body):14.@pl)
[/S]
;================================================================
```

The above trace provides a good illustration to the PPP theory. *reach* is the main predicate of the sentence and is preserved till the end of the analysis.

4 English Analysis

We have been guided by the basic grammatical constructs of the language (Wren and Martin, 1991) to create the system, while the evaluation of the system has been done on actual corpora as described in the next section. Numerous phenomena have been handled. We give a typical and non-trivial case- that of *Gerunds*- to give a flavour of the process. The rest of the phenomena are just mentioned for want of space.

4.1 Gerunds

A Gerund is that form of verb, which ends with *–ing*, and has the force of both noun and a verb. The main problem is to detect when the gerund is functioning as a verb and when as a noun. The situations are so varied that that the solution is necessarily heuristic.

Reading *is my favourite past time*

Reading *books is my favourite past time*

In the first example, *reading* has the force of a noun, while in the second it has force of a noun as well as a verb, as *reading* takes *book* as an object.

Consider the second sentence first. Our strategy to solve this problem is to let the gerund behave as a verb. After the object (*i.e.,* resolving the *object* relationship with *book)* has been accounted for*, book* is deleted from node-list. What remains in the node list thereafter is the same as the first sentence. Now rules are needed to convert this verb into the noun form.

When a sentence starts with a verb in *–ing* form and is followed by an auxiliary verb, then the first verb in –ing form has to be a gerund. This heuristic is enough for us to detect *reading* as a noun. The rule given below serves this purpose.

```
R(SHEAD){VRB,CONT,^N:-VRB,+N,+METHOD::}{VAUX:::}P30;
```

The above rule checks for a sentence starting with a verb in continuous (CONT) form followed by another verb, removes the VRB attribute from the continuous verb and adds attributes N and METHOD to it and then the analysis windows shift to right.

Now all the rules for noun will get applied to *reading* and the analysis for the sentence of the form *"<NP> is my favourite past time"* will proceed as usual.

4.2 Other phenomena

The following phenomena (Wren and Martin 1991) have been handled with very interesting rules and heuristics (Dave et. al. 2002).

a. Nominals (nouns, pronouns, adjectives) in nominative, accusative, possessive and all other cases.

b. Adverbs of time, place, negation, manner, frequency and reason.

c. Prepositions of source, target, restriction, sequence, place, time, purpose.

d. Co-ordinating, subordinating and correlative conjunctions.

e. Infinitives and participles.

f. Adjective, adverb and noun clauses.

g. Pleonastics.

h. Ellipsis.

i. Limited Anaphora.

5 Analysis of Hindi

The rule base that drives the Hindi Analyser (HA) uses strategies different from its English counterpart. This is due to the numerous structural differences between Hindi and English (Tiwari *et. al.* 1987, Gopinathan *et. al.* 1993). But the fundamental mechanism of the system is the same, *i.e.*, it performs morphological, syntactic and semantic analysis synchronously.

The rule base of the HA can be broadly divided into three categories – *morphological rules, composition rules* and *relation resolving rules*. Morphology rules have the highest priority. This is because unless we have the morphed word, we cannot decide upon the part of speech of the word and its relation with the adjacent words. Hindi has a rich morphological structure. Information regarding person, number, tense and gender can be extracted from the morphology of nouns, adjectives and verbs. An exhaustive study of the morphology is done for this purpose and appropriate rules are incorporated into the system. To illustrate the process of Hindi analysis, we consider the following example of a Hindi sentence with an explicit pronoun.

H1. मैंने देखा कि सीता सब्ज़ी खरीद रही है।
mai ne dekhaa ki seetaa sabjee khareed rahee hai
I saw that sita vegetable buying-is

E1. I saw that Sita is buying vegetables.

The processing of this sentence is carried out as follows:

1. The beginning of the clause is marked by the presence of the relative pronoun *ki* (that).
2. The analysis windows right shift till the predicate *dekhaa* is reached.
3. All the relations of the previous nodes with this predicate are resolved. In this case, *mai* (I) being *first person singular* and *animate* pronoun, *agt* relation is produced between *mai ne* and *dekhaa.*
4. The relative pronoun *ki* is now detected and the analysis heads right shift. It combines *ki* with *dekhaa* and adds a dynamic attribute *kiADD* to *dekhaa*.
5. The clause following *ki* is now resolved. The analysis windows right shift till the main predicate of the sentence- *khareed rahee hai-* is reached.
6. It combines the nodes *sabjee* and *khareed rahee hai* with the *obj* relation seeing the *inanimate* attribute of *sabjee.*
7. It then resolves the *agt* relation between *seetaa* and *khareed rahee hai* seeing the *animate* attribute of *seetaa*.
8. At the end of its analysis, its main predicate is retained which in this case is *khareed rahee hai*. Finally the *obj* relation is generated between this verb and *dekhaa*.

Now we describe the various Hindi language phenomena handled by the system. Hindi is a null subject language. This means that it allows the syntactic subject to be absent. For example, the following sentence is valid in Hindi.

H2. जा रहा हूँ।

jaa rahaa hun

going-am

E2. *am going[2]

The system makes the implicit subject explicit in the UNL expressions. The UNL expression produced by the system in this case is:

[S]
agt(go(icl>do).@entry.@present.@progress, I(icl>person))
[/S]

The system can also handle limited amount of anaphora. For example, consider the following sentence:

H3. मेरी ने अपनी किताब जीम को दी है।

meree ne	apanee	kitaab	jeem ko	dee hai
Mary	her	book	Jim-to	given-has

E3. Mary has given her book to Jim.

The corresponding UNL relations generated are:

[S]
pos(book(icl>publication):0C,Mary(icl>person):00)
ben(give(icl>do):0R.@entry.@present.@pred,Jim(icl>person):0J)
obj(give(icl>do):0R.@entry.@present.@pred,book(icl>publication):C)
agt(give(icl>do):0R.@entry.@present.@pred,Mary(icl>person):00)
[/S]

That resolution of the anaphora is apparent from the fact that the UW *she(icl>person)* for *her* is replaced by *Mary(icl>person)* in the *pos* relation.

One of the major differences between Hindi and English is that a single pronoun वह [vah](*he* or *she*) in Hindi is mapped to two pronouns *he* and *she* of English. The gender of the pronoun in Hindi can be known only from the verb morphology. So the system defers the generation of the UW for वह [vah](*he* or *she*) until the verb morphology is resolved. At the end of the analysis, the correct *he(icl>person)* or *she(icl>person)* is produced. For example,

[2] * indicates incorrect grammatical construct

H4. वह शाम को आएगी।

vah shaam ko aaegee

She evening-in will come

E4. She will come in the evening.

The UNL expressions are:

[S]
tim(come(icl>do):0D.@entry.@future,evening(icl>time):05.@def)
agt(come(icl>do):0D.@entry.@future,she(icl>person):00)
[/S]

Hindi uses the word-forms आएगा [aaegaa] and आएगी [aaegee](both meaning *will come*) for the verb आ [*aa*] (come) for a male subject and female subject respectively. Thus, in the above sentence, the verb आएगी [aaegee] causes the UW *she(icl>person)* to be generated for वह [vah](*he* or *she*).

Hindi being a relatively free word-ordered language, the same sentence can be written in more than one way by changing the order of words. For example,

H5. (A) तुम कहाँ जा रहे हो?

tum kahaan jaa rahe ho?

You where going are

(B) कहाँ तुम जा रहे हो?

kahaan tum jaa rahe ho?

where you going-are

(C) कहाँ जा रहे हो तुम?

kahaan jaa rahe ho tum?

where going-are you

E5. Where are you going?

The output in all cases is:

[S]
plc(go(icl>do):07.@entry.@interrogative.@pred.@present.@progress, where(icl>place):00)
agt(go(icl>do):07.@entry.@interrogative.@pred.@present.@progress, you(icl>male):0I)
[/S]

This is achieved as follows. Additional rules are added for each combination of the word types. Also the rules are prioritised such that the right rules are picked up for specific situations. For the sentence H10(A), first the rule for generating *plc* relation between *kahaan* and *jaa rahe ho* is fired, followed by the rule for generating *agt* relation between *tum* and *jaa rahe ho*. In H10(B), first *agt* and then *plc* are resolved. In H10(C), a rule first exchanges the positions of *jaa rahe ho* and *tum*. After that the rules fire as before for setting up the relations. Use is made of the question mark at the end of the sentence.

Hindi allows two types of constructions for adjective clauses– one with explicit clause markers like जो [jo](*who*), जिसकी [jisakee](*whose*), जिसे [jise](*whom*), *etc.* and the other with the वाला [vaalaa](ing) construction. Our analyser can handle both. For example,

H6. पीटर जो लंडन में रहता है वह यहाँ काम करता है।

peeTar jo london mein rahataa hai vah yahaan kaam karataa hai

Peter who London-in stays he here work-do-is

E6. Peter who stays in London works here.

H7. लंडन में रहनेवाला पीटर यहाँ काम करता है।

london mein rahanevaalaa peeTar yahaan kaam karataa hai

London-in staying Peter here work-do-is

E7. Peter who stays in London works here.

The system produces the following UNL relations for both these:

[S]
agt(work(icl>do).@entry.@present, Peter(icl>person))
plc(work(icl>do) .@entry.@present, here)
agt(stay(icl>do) .@present, Peter(icl>person))
plc(stay(icl>do) .@present, London(icl>place))
[/S]

The two incoming arrows into *Peter(icl>person)* provides the clue to the system to correctly identify the adjective clause in each sentence.

6 Evaluation

The reader is urged to see the appendix to have a feel for the complexity of the sentences handled. Table 1 shows the systems statistics of the analyser system in terms of number of rules. As can be observed from the table, control (shift) rules form about 50% of the total number. The reason for this is that many language phenomena are very complex and the system needs to look at a number of words to the right to be able to make correct decisions. The semantic attributes in the lexicon are generated from an evolving inventory of 75 attributes defined for the problem.

The category disambiguation and the elaborate description of the handling of the language phenomena discussed above throw light on some of the capabilities of the system. We have carried out an extensive evaluation of the system, which we now describe. The evaluation results are from techno scientific documents, literary work and the famous *Brown Corpus*.

Type of Rule	Symbol	Number of Rules
Left Composition	+	**178**
Right Composition	-	**91**
Left Modification	<	**703**
Right Modification	>	**596**
Left Shift	L	**152**
Right Shift	R	**2154**
Attribute Assigning	:	**235**
Backtrack	?	**6**
Copy	C	**2**
Syntactic Tree Copy	G	**9**
Left Node Deletion	DL	**0**
Right Node Deletion	DR	**3**
Total Number of Rules		**4131**

Table 1: System Statistics in terms of number of rules

The UNL expressions produced by the system are verified manually, as well as by observing the quality of the Hindi sentences generated using a UNL to Hindi Generator. Currently the system is capable of producing UNL expressions for varied and complex sentences (*vide* appendix).

We have evaluated the system on *Brown Corpus, Medline Text,* agricultural documents, *TREC data* etc., in addition to some literary text. The results of some of these studies is shown in appendix-A.

The *Hindi Analyser* can deal with simple, complex, compound, interrogative as well as imperative sentences. Currently the number of rules in HA is about 3500 and the lexicon size is around 70,000.

7 Related work and comparison

Any system attempting to provide a comprehensive solution to the problem of analysing a natural language has to grapple with the well-known classical problems of *proper noun handling, sense disambiguation, anaphora resolution etc.* While it is difficult to give a complete account of these concerns covering a comprehensive gamut of all the works done, we mention some noteworthy efforts and how they compare with our work.

Our work comes closest to a frame-based interlingua called *Text Meaning Representation* (Boyan and Nirenburg 1992 and also the *Microkosmos* website) in which the meaning of lexical units are represented in terms of mappings into the ontology and/or their contribution to text meaning representations. UNL, however, has better expressiveness as it has wider range of semantic relations and an elegant way of representing clauses in the form of Compound Words (CWs), which is similar to *chunking*.

Some efforts in the areas of classical and shallow parsing of natural language sentences should be compared. Our work is related to *shallow parsing* or *tagging* in the sense that the UNL expressions can be looked upon as *case marker tags* on the words of the sentences. *The Link Parser* is based on link grammar (Slater and Temperley 1998). The major difficulty in link parsing is how to choose the right parse from the set of parses produced by the Link Parser (Perraju 2000). The Parser uses the semantic categories of the WordNet and thus has little flexibility of introducing new semantic attributes in the lexicon to improve the performance of the system. Our system is more flexible as it uses its own lexicon which can be semantically enriched on demand (Verma and Bhattacharyya 2002).

No analyzer of texts can be oblivious to the problem of word sense disambiguation. Be it multiple senses of words, be it multiple parses of sentences, be it multiple users models, the problem of disambiguation is ubiquitous. There are two basic approaches to POS disambiguation, *rule based* (Greene and Rubin 1991) and *probabilistic (e.g., CLAWS tagger,* Leech and Garside 1991). Brill (Brill, E., 1995) has implemented an unsupervised rule based tagger.

Our system can be classified under the rule-based approach. Here the tagging is done by forcing selection of the appropriate entries from the dictionary through rule application and look ahead. The process is more like deterministic parsing (Milne 1988). Our system makes extensive use of *lexical attributes* and *condition windows.* This strategy is used for prepositional phrase disambiguation, where lexical attributes of the nouns are used extensively for disambiguation. The condition windows essentially provide *look-ahead* and *look-back*. Sense disambiguation, however, is a complex problem requiring extensive lexical resources (Yarowski 1992, Yarowski 1995, Bhattacharyya and Narayan 2002, Ramakrishnan and Bhattacharyya 2002). Our system does not attempt sense disambiguation.

Anaphora resolution is a key problem in natural language processing. Our strategy for anaphora resolution makes heavy use of the attributes in the nodes of the condition windows, besides of course the properties of nodes under the analysis windows. Our approach contrasts with approaches based on, for example, *shallow processing* (Carter 1987), *corpus* (Dagan and Itai 90]), *knowledge independence* (Nasukawa 94), *machine learning* (Connolly, Burger and Day 94) and *reasoning based on uncertainty* (Mitkov 95b). Anaphora resolution has remained a challenge with most of the efforts making use of workable heuristics.

8 Conclusion

This paper described a novel approach to language analysis called *predicate preserving parsing*. The synchronous morphological, syntactic and semantic analysis has been used to tackle numerous phenomena of English and Hindi. The system performs very well for many complex phenomena, *e.g.*, *category disambiguation* (*the soldiers went to the totally deserted desert to desert the house in the desert)*. An interesting study has been carried out on the influence of *language divergence* on inter lingua based English Hindi MT (Dave et. al. 2002). The output of the analysis process is tested for its quality and correctness by the reverse process of generation to either the same language or another target language. Results on standard corpora show the promise of the approach.

References

1. Bhattacharyya P. and Narayan Unny, 2002. *Word Sense Disambiguation and Text Similarity Measurement Using WordNet* , Real World Semanic Web Applications, Vipual Kashyap and Leon Shklar (ed.), IOS Press, Amsterdam.

2. Boyan, O. and Nirenburg, S., 1992. Lexicon, Ontology, and Text Meaning. In James Pustejovsky (ed.), *Lexical Semantics and Knowledge Representation*, Heidelberg: Springer Verlag.

3. Brill, E., 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics,* 21(4):543—565.

4. Brown, Peter F., Pietra, Stephen A., Della, Pietra, Della, Vincent J., and Mercer, Robert L., 1991. Word-Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Conference of the Association for Computational Linguistics,* pp. 264-270, Berkeley, CA.

5. Carter D.M., 1987. *Interpreting anaphora in natural language texts,* Chichester: Ellis Horwood, 1987.

6. Church, K., 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing,* pages 136-143, Austin, Texas.

7. Connoly D., Burger J., Day D., 1994. A Machine learning approach to anaphoric reference. In *Proceedings of International Conference on New Methods in Language Processing,* UMIST, Manchester.

8. Dagan I., Itai A., 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of 13th International Conference on Computational Linguistics, COLING'90,* Helsinki.

9. Dave S., Parikh J. and Bhattacharyya P., 2002. *Interlingua Based English Hindi Machine Translation and Language Divergence*, Journal of Machine Translation, Volume 17.

10. EnCo Specification Version 2.1, 2000. UNU/IAS/UNL Centre, Tokyo 150-8304, Japan.

    http://www.unl.ias.unu.edu/unlsys/enco/index.html.

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

12. Gopinathan, S. and Kandaswamy S. (Eds) :1993, *anuvad ki samasayen [Problems of Translation],* Lokbharti Prakashan.

13. Greene, B. B. and Rubin, G. M., 1991. Automatic grammatical tagging of English. Technical Report, Brown University.

14. Hutchins, W. J. and Somers, H. L., 1992. *An Introduction to Machine Translation,* Academic Press.

15. Milne, R., 1988. Lexical Ambiguity Resolution in a Deterministic Parser in Steven. I. Small, Garrison W. Cottrell, Michael K. Tanenhaus (ed.), *Lexical Ambiguity Resolution,* Morgan Kaufman Publishers.

16. Mitkov R., 1995. An uncertainty reasoning approach to anaphora resolution. In *Proceedings of Natural language Pacific Rim Symposium,* Seoul, Korea.

17. Nasukawa T., 1994. Robust method of pronoun resolution using full-text information. In *Proceedings of 15th International Conference on Computational Linguistics COLING'94,* Kyoto, Japan.

18. Neal, J. G., Shapiro, S. C., 1987. *Natural Language Parsing Systems.* Springer-Verlag Publications, New York Berlin Heidelberg.

19. Perraju, M. V. S. S., 2000. Conversion of Natural Language Text into Semantic Net like Structures, Technical report, Department of Computer Science, IIT Bombay.

20. Ramakrishnan G., Prithviraj B.P., A. Deepa, Pushpak Bhattacharyya, and Soumen Chakrabarti, 2004. *Soft Word Sense Disambiguation*, International Conference on Global Wordnet (**GWC 04**), Brno, Czeck Republic.

21. Sanderson, M., 1997. Word Sense Disambiguation and Information Retrieval, PhD Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK.

22. Slater, D. and Temperley, D., 1998. Parsing English with a link grammar. Technical report, School of Computer Science, Carnegie Mellon University.

23. Tiwari Bholanath and Naresh-Kumar: 1987, *Videshi bhashaaon se anuvad ki samasyayen [Problems of translation from various foreign languages],* Prabhat Prakashan.

24. The Universal Networking Language (UNL) specifications version 3.0, 1998. Technical Report, United Nations University, Tokyo.

    http://www.unl.unu.edu/unlsys/unl/unls30.doc

25. Williams, Woods, 1985. What's in a Link: the Foundation for Semantic Networks. In R.J Brachman and J. Levesque (eds.), *Readings in Knowledge Representation*, Morgan Kaufmann Publishers.

26. Verma N. and Bhattacharyya P., 2004. *Automatic Lexicon Generation through Wordnet*, International Conference on Global Wordnet (**GWC 04**), Brno, Czeck Republic, January, 2004 (to be held).

27. Wren, P. C. and Martin, H., 1991. *High School English Grammar and Composition*, S. Chand & Company.

28. Yarowsky, D., 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *In Proceedings of COLING92*, Nantes, France.

29. Yarowsky, D., 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* pages 189--196, Cambridge, MA.

Appendix

We present two sentences from the Brown Corpus for their interestingness. They are from the *br-a01* corpus which is a collection of press reports.

```
;================================UNL============================
[S]
;The progress of science over these last few centuries and the gradual replacement of
Biblical by scientific categories of reality have to a large extent emptied the spirit world of
the entities which previously populated it
{unl}
obj(empty(icl > discharge):4Y.@entry.@present.@complete.@pred,          spirit world(icl >
         imaginary_place):5A.@def)
aoj(empty(icl > discharge):4Y.@entry.@present.@complete.@pred,          :02)
man(empty(icl > discharge):4Y.@entry.@present.@complete.@pred,          to a large
extent(icl>scale):4G)
and:02(replacement(icl > replacing):2N.@entry.@def,      progress(icl >
         advancement):0C.@def)
ben:02(replacement(icl > replacing):2N.@entry.@def,      category(icl > class):2P.@pl)
mod:02(replacement(icl > replacing):2N.@entry.@def,      category(icl > class):3P.@pl)
aoj:02(grading(icl > gradual):2F, replacement(icl > replacing):2N.@entry.@def)
aoj:02(biblical(icl > scriptural):32,          category(icl > class):2P.@pl)
mod:02(category(icl > class):3P.@pl,          reality(icl > world):43)
aoj:02(scientific(icl > rational):3E,          category(icl > class):3P.@pl)
mod:02(progress(icl > advancement):0C.@def,      :01)
mod:02(progress(icl > advancement):0C.@def,      science(icl > skill):0O)
obj:01(over:11.@entry,   century(icl>period):1S.@pl)
mod:01(century(icl >period):1S.@pl,       this:1D.@pl)
mod:01(century(icl > period):1S.@pl, few(icl >small_indefinite_number):1O)
mod:01(few(icl > small_indefinite_number):1O,     last(icl > stopping_point):1J)
mod(spirit world(icl > imaginary_place):5A.@def, entity(icl > something):5U.@def.@pl)
```

```
aoj(populate(icl > dwell):6K.@past.@complete.@pred,      entity(icl >
something):5U.@def.@pl)
man(populate(icl >dwell):6K.@past.@complete.@pred,      previously(icl > initially):69)
obj(populate(icl >dwell):6K.@past.@complete.@pred,      it(icl>thing):6U)
[/S]
;=====================================================
```

This is an interesting sentence. It is to be noted that (i) the elliptical reference to the word *category* (after *biblical)* and (ii) the long distance conjunction between *progress* and *replacement* have been correctly resolved.

```
;======================== UNL ========================
;The Fulton County Grand Jury said on Friday that an investigation of Atlanta's recent primary
election produced no evidence that any irregularities took place
[S]
obj(say(icl>do):0T.@entry.@past.@pred,       :02)
aoj(say(icl>do):0T.@entry.@past.@pred,       Fulton County Grand Jury(icl>group):04.@def)
tim(say(icl>do):0T.@entry.@past.@pred,       friday(icl>day>time):0Y)
aoj:02(produce(icl>happen):2U.@entry.@past.@pred,investigation(icl>inquiry):1F.@indef)
obj:02(produce(icl>happen):2U.@entry.@past.@pred,evidence(icl>information):3B)
aoj:02(no:38,            evidence(icl>information):3B)
aoj:02(:01,   evidence(icl>information):3B)
aoj:01(take place(icl>happen):48.@entry.@past.@pred,           irregularity(icl>misbehavior):3T)
aoj:01(any:3P,           irregularity(icl>misbehavior):3T)
mod:02(investigation(icl>inquiry):1F.@indef, primary election(icl>election):2D)
pos:02(primary election(icl>election):2D, atlanta(icl>state        capital):1W)
aoj:02(recent(icl>past):26,     primary election(icl>election):2D)
[/S]
;====================================================
```

The narrative object for *say* has been capture under the scope **:02.** The phrase *no evidence* has been expressed by considering *no* as an adjecive of *evidence*, though *evidence@not* would have been better.

The following sentence is from a technical manual.

```
;======================== UNL ========================
;If there is an applicable rule then EnCo will add or delete Lexical attributes from these nodes
and create a partial syntactic tree and UNL network according to the type of the rule.
[S]
or:02(delete(icl>erase):1O.@future.@pred,     add(icl>augment):1H.@pred)
nam:02(attribute(icl>property):23,     Lexical:1V)
mod:02(node:2O.@pl,             this:2I)
obj:02(delete(icl>erase):1O.@future.@pred,    attribute(icl>property):23.@pl)
scn:02(delete(icl>erase):1O.@future.@pred,    node:2O.@pl)
aoj:02(delete(icl>erase):1O.@future.@pred,    EnCo:10)
aoj:02(syntactic(icl>grammatical):3F,tree(icl>thing):3P.@indef)
aoj:02(partial(icl>incomplete):37,       tree(icl>thing):3P.@indef)
mod:02(network(icl>thing):42,            UNL:3Y)
and:02(network(icl>thing):42,            tree(icl>thing):3P.@indef)
and:02(create(icl>make):2Y.@entry.@pred.@present, delete(icl>erase):1O.@future.@pred)
obj:02(create(icl>make):2Y.@entry.@pred.@present, network(icl>thing):42)
mod:02(type(icl>kind):4R.@def,         rule:53.@def)
man:02(create(icl>make):2Y.@entry.@pred.@present,             according to:4A)
obj:02(create(icl>make):2Y.@entry.@pred.@present, type(icl>kind):4R.@def)
aoj:01(applicable:0F,            rule:0Q.@entry.@present.@indef)
aoj:01(rule:0Q.@entry.@present.@indef,       there:03)
con(:02.@entry.@pred.@present,       :01)
[/S]
```

```
;====================================================
```

Note that the scopes of various conjunctions are correctly identified (**and** and **or** relations above). While *add* and *delete* are close to each other, *create* is quite far.

Besides the techno-scientific domain, we have also tried the system in the domain of literary works. Here the observation is that the system needs some pre and post editing. Pre-editing involves the controlled use of punctuation and the explicit addition of implicit relative pronouns for certain clausal cases. Post-editing involves correcting the restrictions of some UWs produced by the system. An example of a sentence not handled properly by the system from Wodehouse is:

> *I loosed it down the hatch, and after undergoing the passing discomfort, unavoidable when you drink Jeeves's patent morning revivers, of having the top of the skull fly up to the ceiling and the eyes shoot out of their sockets and rebound from the opposite wall like racquet balls, felt better.*

However, with some obvious pre-editing as shown below the sentence is analysed properly.

I loosed it down the hatch and after undergoing the passing discomfort which is unavoidable when you drink Jeeves's patent morning revivers, of having that the top of the skull fly up to the ceiling and the eyes shoot out of their sockets and rebound from the opposite wall like racquet balls, felt better.