

UNIVERSAL NETWORKING LANGUAGE - A TOOL FOR LANGUAGE-INDEPENDENT SEMANTICS ?

*Amitabha Mukerjee, Achla M Raina, Kumar Kapil, Pankaj Goyal, Pushpraj Shukla
Indian Institute of Technology, Kanpur, India
{amit, achla, kapil, pankajgo, praj}@iitk.ac.in*

Abstract- *Given source text in several languages, can one answer queries in some other language, without translating any of the sources into the language of the questioner? While this task seems extremely difficult at first sight, it is possible that the ongoing UN sponsored Universal Networking Language (UNL) proposal may hold some clues towards achieving this distant dream. In this paper we present a partially implemented solution which shows how UNL, though not designed with this as the primary objective, can be used as the predicate knowledge base on which inferences can be performed. Semantic processing is demonstrated by Question Answering. In our system as of now, both the text corpus and the questions are in English, but if UNL can deliver on its promise of a single homogeneous language-independent encoding, then it should be possible to achieve question answering and other semantic tasks in any language.*

SEMANTICS MODELS AND UNL

Many organizations worldwide are grappling with problems like the following: Given source text in several European languages, would it be possible to demonstrate semantic understanding in some other language (like Hindi) without explicitly translating any of the sources into the language of the questioner? This is, of course, an extremely difficult task, perhaps even an impossibly difficult task. We trust the reader will realize that this paper is merely a very preliminary investigation as indicated by the hesitant “?” at the end of the paper’s title. The key insight driving this research is the realization that if there is a mechanism for mapping any language into a uniform language-independent predicate structure, then it would constitute an important tool in this direction. While no system worldwide is anywhere near succeeding in this effort, the ongoing work on Universal Networking Language (UNL) [2] appears to hold the highest promise in terms of delivering on this dream.

UNL was developed as a universal knowledge-encoding mechanism, and is being primarily driven by the needs of the MT community. UNL provides for a uniform concept vocabulary (called “universal words” or UW’s – the same concept in any language results in the same UW, which is written out using English orthography). These UW’s are connected by a small set of about thirty-eight binary

relations to obtain a set of predicate expressions that can encode the linguistic content of any sentence in any language of the world. One of the philosophical issues of course, is that the same concept, as expressed in different languages, does not define an identical chunk of conceptual space and at best, the UW’s are approximations to the overlapping part of this concept. Despite such philosophical indulgences, a number of groups around the world are working on constructing a UNL KB (a knowledge structure linking the concepts underlying the UW’s in terms of the probability of certain relations holding between them), and on constructing converters (from NL to UNL) and deconverters (from UNL to NL) for several languages across the world. In this latter sense, the UNL may be thought to be an inter lingua, but UNL has a number of other features that make it better suited for semantic inference than most other interlinguas. In particular, the following features of UNL motivate this work:

1. The set of Universal Words with well defined universal interpretations,
2. a small, simple predicate structure with only binary predicates,
3. a knowledge base connecting the UW’s as a weighted graph of relations.
4. ontological information that is built-in to the UWs (eg. cholera(icl<disease) characterizes cholera as a type of disease).
5. The world wide effort in developing mechanisms for converting language into UNL and vice versa.
6. The dream of language independent semantic analysis.

Even aside from the language independence claim, there are merits to using a coherent labeling structure as provided by the UW’s. Models for semantics require a basic set of predicates into which sentences from Natural Language would be mapped. All such efforts, e.g. CYC[10] have been plagued by considerable divergence in semantic analysis. By removing multiple models of reflecting the same concept at source, UW’s help this objective significantly. It may be argued that other tools such as WordNet [11] provide a richer ontology and lexical knowledge for this task, but they

do not provide the predicate structure, or the en/de - converting tools of UNL.

PRESENT WORK

This work makes two major claims:

- that a substantial amount of logical inference is possible on the UNL representation of language as UNL expressions,
- that for the question answering task in particular, given that converters to UNL and deconverters from UNL are indeed available, the only task remaining to achieve such an objective is to construct the question to answer-template UNL mapping for the target language.

Some of these goals are clearly far in the future and even if it happens that some aspects of the UNL experiment may not quite succeed, it is still likely that the effort would lead to insights applicable to any other model for language-independent semantics.

In our implementation we demonstrate both inference, and question answering - though at this point, both of these operate on English alone. The Q&A is achieved using a content-level HPSG lexicon tagged with the appropriate UNL relations. In earlier work we have used the same HPSG structure for looking at English, Hindi and also codemixed bilingual structures, and in future, we hope to demonstrate the Q&A aspects in Hindi with the source text in English itself.

In practically implementing this interface, we have to deal with an actual corpus written in some language, and also a procedure for testing the degree of success in modeling the semantics. In this case, we have chosen an English-based corpus for safe drinking water and as a test procedure we have developed an English based question and answer system, in the course of which we have also developed a lexicon of transformational rules for a subset of English questions. Since semantic modeling requires models of a body of "commonsense" knowledge and associated pragmatic rules, which in this instance need to be created manually, we have restricted ourselves to a limited domain - that of drinking water.

The Question Answering module comprises traditional modules such as a syntactic parser, logical representation of the text(UNL), built in ontologies (UNLKB), inference engine, question processing, document retrieval, answer extraction and others [1].

From Natural Language to UNL

The corpus for the present work was built from documents obtained from the official websites of EPA and WHO [12]. Since the converters mapping NL corpus into a UNL

document are not yet available [2], the mapping was done manually. To build the corpus KB, we marginally edited the source document, e.g. dropping the phrase "just like man-made chemicals" in the source sentence "At certain levels, minerals, just like man-made chemicals, are considered contaminants that can make water unpalatable or even unsafe" because we could not find a clear definition for 'just like' phrases in the UNL Specifications[3].

The resulting corpus was annotated and processed to generate the UNL document(the UNL expression for the corpus). The manual annotation of the corpus was done making use of a format specified by UNL [3]. For example, the NL corpus sentence, *At some level, minerals are considered contaminants that can make water unpalatable*, is annotated as follows:

```
<c>At some{<qua,>n} level.n.@pl.@entry</c>{<man,>p}
mineral.@pl{<gol,>p} are consider.p
contaminant{1}.@pl{<obj,>p} that{<1>}{<agt,>p} can
make.p.@possible water{<obj,>p} <c>unpalatable{<or,>p} or
even{<man,>p} unsafe.p.@entry</c>{<gol,>p}.
```

The corpus sentence in its annotated form is input to the UNL parser to generate the UNL parsed graph, represented as a list of relations, given below:

```
unl
obj(consider(icl>think(agt>volitional thing,gol>uw,obj>thing)):25,
contaminant:2G.@pl)
man(consider(icl>think(agt>volitional
thing,gol>uw,obj>thing)):25,
:01)
gol(consider(icl>think(agt>volitional thing,gol>uw,obj>thing)):25,
mineral(icl>matter):1G.@pl)
qua:01(level:0K.@entry.@pl, some(mod<thing):06)

/unl

unl
gol(make(agt>thing,obj>thing,src>thing):3U.@possible,:02)
obj(make(agt>thing,obj>thing,src>thing):3U.@possible,
water(icl>liquid):4B)
agt(make(agt>thing,obj>thing,src>thing):3U.@possible,contamina
nt:2G)
or:02(unsafe(aoj>thing):5U.@entry,unpalatable(aoj>thing):4T)
man:02(unsafe(aoj>thing):5U.@entry,even:5G)

/unl
```

INFERENCE ENGINE

Although UNL structure provides a first order logic encoding of natural language, it is not designed for making semantic inferences, and an inference engine needs to be built for this purpose. In addition, world knowledge about

the domain is needed to provide context information which would be commonly known to the human reader but is not available from the text itself. For the purpose of the Question Answering system, the inference engine also provides a set of inferred facts which can be eventually matched with a pseudo-UNL form of the natural language Query in order to obtain an answer.

The domain used here is that of “drinkable water”. The UNL form of the input text consists of a set of UNL expressions, each of which is a binary predicate corresponding to one of the UNL relations. The arguments to these predicates are Universal Words, possibly modified by one or more attributes.

Pragmatic knowledge

A set of manually created rules encode the pragmatic knowledge in the system. These include facts such as the following:

- Water is essential for human life.
- Communicable diseases are caused through physical contact.
- Water-borne diseases are communicable.

The last rule would have a UNL structure as follows in the pragmatic rule base:

```
aoj(communicable(aoj>thing),water-borne
disease(icl>disease).@pl.@entry)
```

Semantic Equivalence

The same information may be expressed in very different ways:

1. Safe water can be obtained through boiling and distillation.
2. We can obtain safe water through boiling and distillation.
3. We can get safe water through boiling and distillation.
4. The methods for making safe water are boiling and distillation.
5. One can make safe water by boiling or distillation.
6. While distilling results in pure water, for practical purposes, boiling is sufficient to make water safe for drinking.

Fortunately, a part of this problem (e.g active vs passive voice) is resolved by the UNL encoding process – thus (1) and (2) will result in the same UNL structure:

```
agt1(get2(agt>thing,obj>thing,src>thing).@possible,we);
obj(get(agt>thing,obj>thing,src>thing),water(icl>liquid));
mod(water(icl>liquid),safe(mod<thing));
```

© Convergences '03

International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies

```
man(get(agt>thing,obj>thing,src>thing),
through(icl>how(obj>thing)));
obj(through(icl>how(obj>thing)),:01);
and :01(boiling(icl>act),distillation(icl>act));
```

Even (3) which uses the word “get” which is used here in the same sense as “obtain” results in the same universal word

```
get(agt>thing,obj>thing,src>thing)
```

and thus result in the same UNL structure. However sentences (4 and 5) use “make” which has a different UW, and these are handled in the inference engine by using rules for unifying similar UWs when used in the context of water. Very wide variations such as (6), which requires added pragmatic knowledge such as “pure water is safe”, and also results in a set of two conjunctive UNL expressions (one for the “while” clause, and the other for the main clause) can be handled but since the set of such constructs is very large, they are not handled in the current version.

First order Inference Rules

These rules implement the First Order Logic in order to obtain new inferences. For example, given the facts *Water-borne diseases are caused by ingestion of contaminated water.* and *cholera is a water-borne disease.*, one may infer that *cholera is caused by ingestion of contaminated water.*

A meta-rule for this situation, incorporated as part of the inference rulebase is that, given:

```
agt(cause(icl>abstract thing).@entry:1);
obj(cause(icl>abstract thing).@entry:2);
nam(2:3);
```

which says that variable 1 causes 2, and 3 is a type of 2. Given this set of UNL relations, the meta-rule says that one can infer:

```
agt(cause(icl>abstract thing).@entry:,1:);
obj(cause(icl>abstract thing).@entry:,3:);
```

i.e. 3 is caused by 1. The current system is designed to be tested only on a simple Question and Answer mechanism. We use single-tiered inferences, and construct a complete set of all possible inferences that can be made from the given text and the pragmatic rules. This is the final UNL knowledge base which is to be matched with a UNL form of the question to obtain the answer.

December 2 - 6, 2003, Alexandria, EGYPT

THE QUESTION AND ANSWER MODULE

We use a structure matching approach to search for the answer to a question. This is done by building an answer template that represents the form of the potential answer corresponding to the question. This template is input to an HPSG Parser [7, 8] which outputs a pseudo UNL expression corresponding to the question as well as the answer template. The pseudo UNL expression is subsequently subject to a structure matching with the UNL document (the corpus Knowledge Base)

Question Processing

We generate an answer template that represents the form of the answer corresponding to a question with the "exact answer" slot filled in with an unknown variable "X". The existing literature on Q/A systems suggests several ways of building the template such as generic extraction using shallow parsing rules[4]. In this work we use a set of transformational rules to arrive at the answer template. The transformation from the question to the answer template is governed by a rule base with approximately 50 rules which range over various "wh" and other question formats, such as the "yes/no" question. The rules introduce the variable "X" at the location of the keyword or the key phrase in the answer pattern.

For example, the rule, *how:aux:1:V(ppl) > 1:aux:V(ppl):by:X*, works upon a question such as *How is water contaminated?* which is transformed to its corresponding answer template with the variable "X" - *Water is contaminated by X*.

Taking another example, a rule of the form, *what:does:1:V(base) > 1:V(pres):X*, maps the question, *What does skin or eye contact with water cause?* into the answer template *Skin or eye contact with water causes X*.

The Pseudo-UNL Enconverter

The answer template is converted into a pseudo UNL representation by a parser [8] which operates on a lexicon specifying the semantic selection (as against the categorical selection) properties of heads. Semantic relation attributes are used instead of syntactic subcat features since the parsed answer form needs to be unified with a database that is in the UNL format, i.e. the UNL Document. The UNL structure uses relations that are defined in terms of semantic features such as agency, place, etc. Therefore, these relations need to be identified in the parsed answer form for structure matching to be possible. To take an example of a lexical entry stating the said semantic feature information:

```
<make>@V(base){agt||obj|~gol}
```

In the entry for the verb "make" above, the description "{agt||obj|~gol}" captures the fact that the verb phrase headed by the verb "make" takes the form of an Agent followed by the verb itself and then an Object and an optional Goal". Note that agt, obj and gol are all UNL relations.

The Nominal heads which take on the roles agent, object and goal are entered in the lexicon as follows:

```
<impurities>@agt(pl){~qua|~mod||~plc}
<water>@obj(sg){~qua|~mod||~plc}
<unpalatable>@gol{!}
```

The HPSG parser reads the lexicon and states relations as given in lexical entries. From the parsed tree thus obtained, we can get the relation between two nodes, which would essentially be the label attached with the child node. To take an example of how the pseudo-enconverter works, given the question, *What makes water unpalatable?*, we generate an answer template, *X makes water unpalatable*, with the transformational rule,

what:V:1 > X:V:1

The parsed output is as follows:

```
(( Xagt(sg)) makesV(base) (waterobj(sg)) (unpalatablegol) )
+X makes water unpalatable
+X_agt(sg)
+-makes water unpalatable
+-makes_V(base)
+-water_obj(sg)
+-unpalatable_gol
```

The list of relations produced is -

```
agt(makes,X)
obj(makes,water)
gol(make,unpalatable)
```

Similarly, for the question, *How is cholera caused?*, we generate an answer template, *Cholera is caused by X*, with the transformational rule,

how:aux:1:V(ppl) > 1:aux:V(ppl):by:X

The parsed output is as follows:

```
(( choleraobj(sg) ) (isaux) causedV(ppl) ( by ) ( Xagt(sg) ) )
+-cholera is caused by X
+-cholera_obj(sg)
+-is_aux
+-caused by X
+-caused_V(ppl)
+-by_by
+-X_agt(sg)
```

The list of relations produced is -
 obj(caused,cholera)
 aux(caused,is)
 by(caused,by)
 agt(caused,X)

In this case the output is filtered to retain the UNL relations (semantic relations) only i.e.
 obj(caused,cholera)
 agt(caused,X)

Answer extraction from UNL

Given the answer form of the question in pseudo-UNL format, it has to be matched with the final UNL knowledge base to see if an answer can be provided. First, each sentence in the knowledge base (as generated in section 2) is converted into an UNL graph, with two arguments as nodes, connected by a link with the label of the relation. Next, we convert the psuedo-UNL answer template as described in section 3.1 into the UNL graph. For example, given the question *How is Colera caused?*, one obtains the "Query UNL Graph" as in Figure 1.

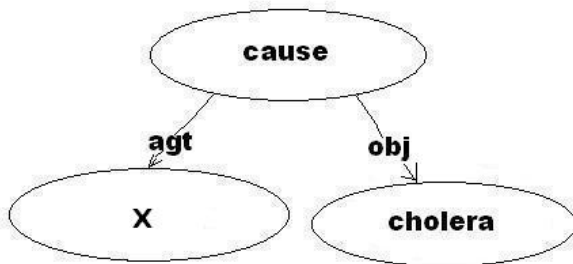


Figure 1: UNL graph for the question "How is cholera caused?"

Finding an answer involves matching this Query UNL graph with a UNL sub graph from the knowledge base, If one of the nodes in the query has a variable X then the match returns the value of this variable. If there is no variable then the match returns T. If no match is found, the system returns F.

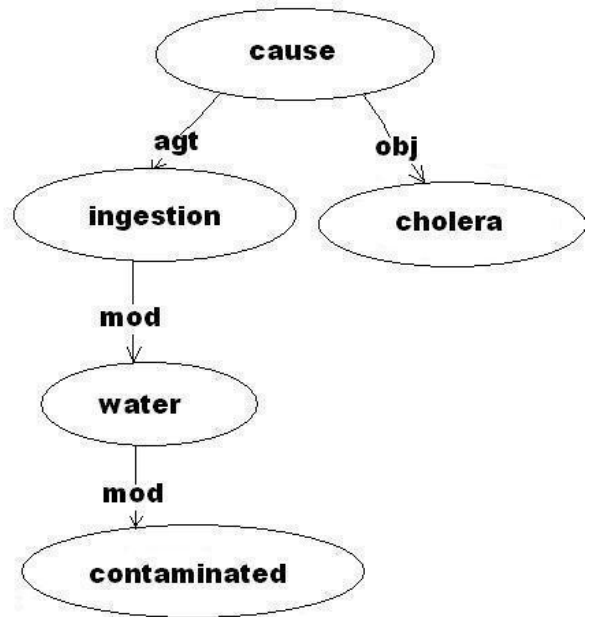


Figure 2: UNL graph for the fact "Cholera is caused by ingestion of contaminated water."

For the above example, there is an inferred fact in the knowledge base for which the corresponding graph is shown in Figure 2. After sub graph matching, X is bound to the left-child of the node "cause" in Figure 2.

In fully implemented UNL situations, this graph can now be pasted, into the answer template and passed to a deconverter which would then generate the full answer sentence. In our case, since no deconverter is being used, we label the sub graphs in the knowledge base with English strings, which are then used in the answer generation process to obtain answers as English sentences. Note that if the deconverter is for a different language, then the answers can also be generated in that language. In this case, the resulting answer is Cholera is caused by ingestion of contaminated water. Similarly, the question "What makes water unpalatable?" results in the graph as shown in Figure 3 below,

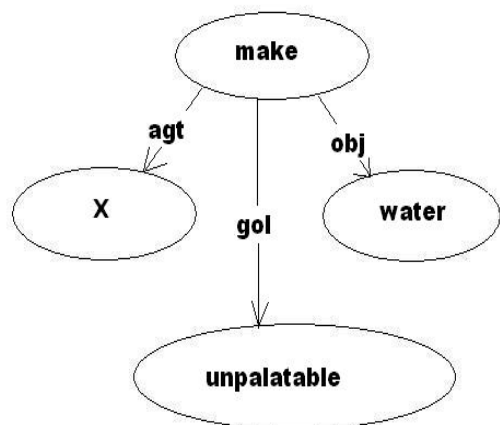


Figure 3: UNL graph for the question "What makes water unpalatable?"

and after matching, results in the answer: "Contaminants make water unpalatable."

CONCLUSION

This work takes a structure intended to represent the structure of a source language and convert it into other languages, and uses it as a query system that can answer questions based on textual databases, possibly in other languages. This is clearly only the first step – a lot more needs to be done to validate the feasibility of this process. A number of important issues remain. While the UNL structure is a First Order Predicate form, there are remarkable differences with normal logical models. For one, UNL structures do not provide for an implication connective, and also use the disjunction relation "or" rather sparingly. A rigorous mapping to more traditional logical structures is needed for more extensive UNL based logical inferencing. Efforts are on in this direction.

Also, the manual process of designing the pragmatic knowledge-base is expensive – it needs to be seen if further synergies can be gained by unifying this effort with parts of the UNL KB. Despite these shortcomings, we hope the present work will provide a start towards this difficult yet important problem. The Q/A module reported here can be successfully extended to other languages without any basic changes in the system design. A UNL-based Q/A system for Hindi, which can work on the Water domain, is expected to be implemented shortly.

ACKNOWLEDGMENT

We are thankful to the UNDL Foundation, Geneva for the Universal Parser, which was used to generate the UNL expressions from the manually annotated UNL document.

REFERENCES

- [1] Moldovan, D., C. Clark, S. Harabagiu and S. Maiorana, "COGEX: A Logic Prover for Question Answering", *Proceedings of HLT-NAACL 2003*, Main Papers, Edmonton, 2003, 87-93.
- [2] "<http://www.undl.org>"
- [3] "UNL Specifications", "<http://www.unlc.undl.org/unlsys/>"
- [4] Diekema, A.R., J. Chen, N. McCracken, N.E. Ozgencil, M.D. Taet, O. Yilmazel, and E.D. Liddy, "Question Answering: CNLP at the TREC-2002 Question Answering Track", <http://www.cnlp.org>
- [5] Robertson, S. and S. Walker, "Okapo/Keebow at TREC-8", *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, Gaithersburg, Maryland, 17-19 November, 1999, 151-162.
- [6] Moldovan, D., M. Pasca, S. Harabagiu and M. Surdeanu, "Performance Issues and Error Analysis in an Open-domain Question-Answering System", *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July, 2002, 33-40.
- [7] Sharma, D., K. Vikram, M.R. Mital, A. Mukerjee, A.M. Raina, "Saarthaka - an Integrated Discourse Semantic Model for Bilingual Corpora", *Online Proceedings of the International Conference on Universal Knowledge and Language*, Goa, India, 25-28 Nov, 2002.
- [8] Sharma, D., K. Vikram, M.R. Mital, A. Mukerjee, A.M. Raina, "Saarthaka - A generalized HPSG parser for English and Hindi", *Recent Advances in Natural Language Processing - Proceedings ICON-2002*, Mumbai, India, 18-21 Dec, 2002.
- [9] Hong, M. and O. Strieter, "Overcoming the Language Barriers in the Web: The UNL- Approach".
- [10] "<http://www.cyc.com/tech.html/cycl>".
- [11] Fellbaum, C., "WordNet: An Electronic Lexical Database", MIT Press, 1998.
- [12] Documents on Water Sanitation and Hygiene "<http://www.who.int/inf-fs/en/fact112.html>", "<http://www.epa.gov/safewater/dwh/contams.html>".