

Using an Interlingua for Document Knowledge Representation

Jesús Cardeñosa

Dpto. Inteligencia Artificial
Universidad Politécnica de Madrid
carde@opera.dia.fi.upm.es

Carolina Gallardo

Dpto. Inteligencia Artificial
Universidad Politécnica de Madrid
carolina@opera.dia.fi.upm.es

Luis Iraola

Dpto. Inteligencia Artificial
Universidad Politécnica de Madrid
luis@opera.dia.fi.upm.es

Abstract

In this paper, the authors advocate in favor of using an interlingua for representing the knowledge contained in text documents. The advocated interlingua, UNL, was designed by the United Nations University to support a language independent textual representation to overcome linguistic barriers in Internet. This paper describes the main features of UNL and presents the application of this interlingua as a Document Knowledge Representation language. This approach is described through the applications developed in two international projects: HEREIN-II (IST-2000-29355) and AgroExplorer.

Keywords: Document knowledge representation, Interlinguas, UNL.

1 Introduction

Recent trends in Document Knowledge Representation (DKR) vary from word sense disambiguation using word sense hierarchies such as Wordnet [1], to shallow syntactic parsing [2], dependency based representations [3] and ontologies based on description logics [4]. Whatever the approach, two dimensions are identified when representing terms and the knowledge involved in a document: the microcontext and the macrocontext (according to the terminology of [3]), being the microcontext the contextual relations in a text and being the macrocontext the lexical and conceptual relations of the sense of a word.

DKR relates to both natural language processing issues and meaning representation in a conceptual,

non linguistic way. A conceptual representation can be considered an interlingual representation of meaning. However, existent conceptual representations are of limited use when employed in tasks such as machine translation, multilingual text generation or information extraction from web documents.

Artificial Intelligence has sought other ways of representing linguistically expressed knowledge, namely, the interlinguas created in the field of Machine Translation (MT). Interlingual MT tried to overcome the exponential explosion of transfer-based MT systems as the number of languages grows. Although interlinguas come from a different research field, there are several facts that support the idea of using an interlingua for representing the knowledge expressed in natural language:

1. Interlinguas deal with the representation of meaning, the most abstract and the deepest level of linguistic analysis. And the interlingual approach attempts to find a meaning representation common to many (ideally to all) natural languages, a representation that leaves aside 'surface' details and unveils a common structure.
2. An interlingua is just another language in the sense that it is autonomous and thus its components need to be defined: vocabulary and relations mainly.
3. Senses and not words are usually the semantic atoms of interlinguas.
4. Thematic and functional relations are established among the semantic atoms of the interlingua, being semantic in nature, allowing for universality and depth of abstraction and analysis.

As can be seen, an Interlingua subsumes the knowledge representation dimensions present in current DKRs.

However, although interlinguas may very well provide the DKR required both for MT and multilingual text generation as well as for large scale knowledge representation tasks, there are some obstacles in the design and further use of an interlingua:

- For multilingual generation and MT purposes, interlinguas are so close to the knowledge level that text generation is hindered by the lack of surface information.
- A proper design of an interlingua is highly complex, since it has been proved almost unfeasible to find a suitable way to represent word meanings that is at the same time a) able to accommodate a wide variety of natural languages, b) easy to grasp and use, c) precise and unambiguous and d) expressive enough to capture subtleties of word meanings expressed in natural languages.

In principle, the first obstacle should not affect to DKR; it would be even desirable, since the more distant the representation from linguistic surface forms, the closer to knowledge it may be. However, as the distance from linguistic surface form grows, the more difficult is to automate the process of converting a natural language text to the interlingual representation. This must be a serious concern in the processing of documents coming from widely different and complex domains such as the Internet, technical document bases, etc.

2 Classical interlingual approaches

Developed within the MT field, classical interlinguas include ATLAS [5] or PIVOT [6]. These interlinguas are paradigmatic of the dominant approach to interlinguas as they are designed as a general domain system for the greater number of languages.

Interlingua-based MT systems did not meet the expectations they created, mainly due to the linguistic problems posed by their insufficiency to express surface phenomena and an incomplete and unsatisfactory account of lexical meaning. However, the development of interlinguas continued and classical interlinguas evolved into the so-called Knowledge Based Machine Translation Systems. Under this label are included the KANT interlingua [7] and the Text Meaning Representations of the Mikrokosmos systems [8]. These interlinguas highlight the knowledge representation dimension of

the interlingua as well as the linguistic aspects, adopting an ontological and frame-based approach for the definition of the concepts. However, the burden of such an intense and detailed knowledge based conceptual modeling can only be afforded in specific domains and for a limited number of language pairs.

Other interlingual devices such as Lexical Conceptual Structures (LCS) [9] are based on sophisticated lexical semantics analysis oriented by linguistic theories [10]. LCS representations are based on a limited number of primitive concepts that serve as building blocks for the definition of all remaining concepts. This approach is well suited for semantic inference, but at the expense of limiting the capabilities of representing the lexical richness present in natural languages.

In general, these interlinguas are hindered by the fact that they are restricted to specific domains. In addition to this, they require a substantial work for building up a conceptual base. The use of semantic primitives may be justified for inference purposes but its actual design and application in a multilingual (or simply in a NLP environment) is difficult and poses more problems than it solves.

In the next section, we will present another attempt to define an interlingua that on the one hand produces a representation of document's content that removes away the details of the source language so it qualifies as a language independent representation while on the other hand keeps enough linguistic information for making feasible text generation in a multilingual environment that includes more than a dozen languages.

3 The Universal Networking Language

During the nineties, the University of the United Nations developed the Universal Networking Language (UNL), a language for the representation of contents in a language independent way, with the purpose of overcoming the linguistic barriers in Internet. UNL has been tested and proved as a language tractable by computing systems, since its expressions can be automatically transformed into those of any natural language by means of a generation processes that follows its specifications [11].

The UNL is composed of three main elements: Universal words, Relations and Attributes. Formally,

a UNL expression can be viewed as a semantic net, whose nodes are Universal words, linked by arcs labeled with UNL relations. Universal Words are modified by the so-called attributes. The language is formally defined in its specifications.

3.1 Universal Words

Universal words (UWs) constitute the vocabulary of the interlingua. To be able to express any concept occurring in a natural language, UNL uses English headwords modified by semantic restrictions that eliminate their semantic ambiguity. If there is no English word suitable to express the concept, UNL allows for the use of words coming from other languages. In this way, the interlingua gets its expressive richness from natural language lexicons but without their lexical ambiguity.

For example, the verb “land” in English has several senses and different predicate frames. The corresponding UW for one of the different senses of this verb in UNL could be:

The plane (agt) landed at the Geneva airport (plc)

The UW for “land” in this sentence would be: `land(icl>do, src>air, agt.@A>thing, plc.@B>thing)`, where “`icl>do, src>air`” are the semantic restrictions that select the intended sense of “land” and “`agt.@A>thing, plc.@B>thing`” presents the argument structure of this predicate.

Although this method is far from perfect, it shows some advantages:

1. There is an agreed and normalized way to define UWs among UNL developers of different language modules and how these UW should be interpreted.
2. It is devoid of the ambiguity inherent to natural language vocabulary.
3. It constitutes the pivot to connect the lexicons of different natural languages.

A first reproach that could be done to this interlingual vocabulary is its anglo-centered vision, which may increase the problem of lexical mismatches among languages. However, this system permits and guarantees expressivity and domain independency. Furthermore, there is another practical advantage derived as a side-effect from the apparent dependency of English: the relations between different UWs are quite similar to those of Wordnet, thus the development of a large UW system can benefit from the lexicographic work

done for creating the Wordnet synset hierarchy as shown in [12]. Many developments in the areas of Information Retrieval and Extraction using Wordnet as well as the usefulness of Wordnet for knowledge extraction tasks can be applied to the UW system, since both share a common foundation. For a more comprehensive discussion of the UW system, see [13].

3.2 Relations

UNL includes a group of 41 basic semantic relations that allows the definition of any possible semantic relation among two concepts. They include argumentative (agent, object, goal), circumstantial (purpose, time, place), logic (conjunction, and disjunction) relations, etc. The UNL specifications provide definitions in natural language of the intended meaning of these semantic relations and establish the contexts in which these relations may apply, such as the kind (semantic class) of the UWs appearing as origin and destination of any relation.

For example, an agent relation can link an action (as opposed to an event or a process) and a volitional agent (as opposed to a property or a substance). This characterization of concepts implies a top “ontology” or “taxonomy” similar to Wordnet’s, whose main purpose is validating the correct application of conceptual relations.

3.3 Attributes

Attributes express several kinds of semantic information that usually modify the predication described by the net of UWs linked with relations. This information includes time and aspect of the event, modality of the predication, reference of the entities mentioned, number and/or gender, etc. For example, in the sentence “The boy eats potatoes in the kitchen”, attributes are needed to express plurality in the object (“potatoes”), definite reference of both the agent (“boy”) and the place (“kitchen”) and finally a special attribute denoting which UW is the head of the whole expression (the entry node). The textual representation of the UNL graph corresponding to this sentence is as follows:

```
agt(eat(icl>do).@entry, boy(icl>person).@def)
obj(eat(icl>do).@entry, potato(icl>food).@pl )
plc(eat(icl>do).@entry, kitchen(icl>facilities).@def)
```

By means of these three components, UNL clearly differentiates between propositional and contextual meaning of utterances: the subset of the language UW + Conceptual Relations defines the

propositional part of a given text, the addition of attributes adds the necessary information related to the participants of the linguistic act and to context.

UNL is accompanied by software tools that facilitate the task of converting a natural language text into UNL and perform automatic generation from UNL texts into natural language.

4 UNL as a language for Document Knowledge Representation

The initial and declared purpose of UNL was to support the exchange of textual information in multilingual environments. However, its design allows for applications that do not revolve around language, that is, UNL could serve as a support for knowledge representation in generic domains. When there is a need to represent knowledge in a domain-independent way, researches turn back to natural language to explore the semantic atoms used by natural languages for expressing knowledge. UNL follows this philosophy, since it provides an interlingual analysis of natural language semantics. UNL can be proposed as a firm document knowledge representation language because:

1. The set of necessary relations existing between concepts is already standardized. Although some of these conceptual relations have a strong linguistic basis other relation groups such as the logical (conjunction, disjunction), temporal, spatial and causative (condition, instrument, method) relations have been widely employed in semantic analysis as well as in knowledge representation.
2. Similarly, the set of attributes that modify concepts and relations is fixed and well-defined, guaranteeing a precise definition of contextual information. Thus, UNL provides mechanisms to clear-cut propositional from contextual meaning.
3. The semantic atoms (UWs) are not concepts but word senses, extracted mainly from the English lexicon for convenience reasons (most if not all languages are provided with bilingual dictionaries, to and from English) and organized according to hierarchical relations.
4. UNL syntax and semantics are formally defined.

The potential of UNL as a document knowledge representation together with its multilingual capabilities are best exemplified in two projects: Herein and AgroExplorer.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rapport (View Source for full doctype...)>
<rapport id="1.3" pays="ES" langue="unl">
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rapport (View Source for full doctype...)>
<stheme id="1.3" contenu="COMPLET">
<para>
  agt(manage(icl>do).@entry.@future,
    Turespaña(iof>institution))
  obj(manage(icl>do).@entry.@future,
    restoration(icl>activity).@def)
  obj(restoration(icl>activity).@def,
    palace(icl>building).@def)
  mod(palace(icl>building).@def, royal(mod<thing))
  plc(palace(icl>building).@def, Madrid(iof>city))
  mod(term(icl>time), short(mod<thing))
  and(short(mod<thing), long(mod<thing))
</para>
```

Figure 2: Embedding UNL into XML documents

4.1 The Herein Project

The Herein project (IST-2000-29355) (HEREIN, 2003) aims at the creation of an Internet-based facility for improving cultural heritage management methods at the European level. Among the main tasks of the project, participant countries must compose a report providing detailed information about all aspects regarding cultural heritage.

Documents on cultural heritage from each country were provided in XML format. Figure 1 shows a fragment of a typical report in original Spanish language.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rapport (View Source for full doctype...)>
<rapport id="1.3" pays="ES" langue="es">
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rapport (View Source for full doctype...)>
<stheme id="1.3" contenu="COMPLET">
<para>
  La restauración del Palacio Real de Madrid será
  gestionada por Turespaña </para>
```

Figure 1: Fragment of the Spanish report

The entire report of the Spanish cultural heritage was codified into UNL, and the UNL representation has been then embedded into the XML structure common to all reports, as if UNL were another “natural language”. Figure 2 shows the UNL representation of one of the sentences of the previous fragment and how it has been embedded into the common XML structure.

The integration of UNL into the Herein system, [14] is illustrated in figure 3. An original XML document about Spanish heritage is the input to an UNL editor,

once the XML tags have been removed and the contents extracted. The UNL editor is a tool where a user can produce UNL representations from Spanish sentences in a semi-automatic way using Spanish-UNL dictionaries and syntactic and semantic UNL analyzers. The output of the UNL editor is a UNL representation in which no XML tagging is preserved. In the project, available English and Russian generators produce the Spanish contents in those two languages. The final step is the “XMLization” of these plain documents according to the DTD adopted in the Herein system.

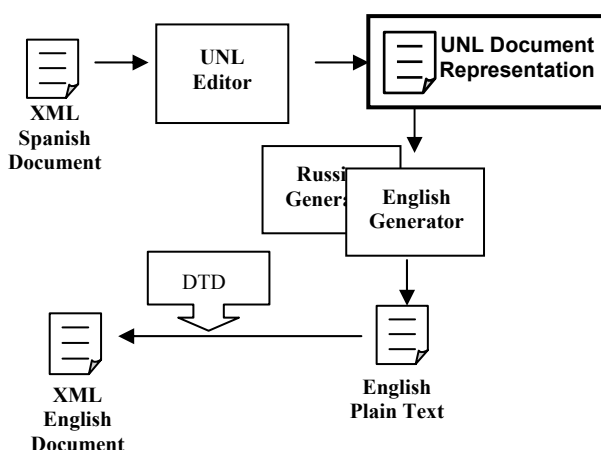


Figure 3. Integration of UNL into Herein

The application of UNL in the Herein proves that UNL can be employed as a general-purpose, language independent document representation formalism. It also opens the way for using UNL representations as truly DKRs. UNL representations allow us to exploit semantic relations among the concepts present in the document. Thus, the document has been enriched with a conceptual dimension that makes possible knowledge-intensive applications such as search systems capable of carrying out reasoning processes in order to produce intelligent answers to natural language queries.

As an illustration of the potential of UNL for such applications, let us consider one of the sentences contained in the previous fragment:

<para> La restauración del Palacio Real de Madrid será gestionada por Turespaña. </para>

(The restoration of the Royal Palace of Madrid will be managed by Turespaña)

As shown in fig. 2, its UNL representation is:

```

[S]
agt (manage(icl>do).@entry.@future,
Turespaña(iof>institution))
obj (manage(icl>do).@entry.@future,
restoration(icl>activity).@def)
obj (restoration(icl>activity).@def,
palace(icl>building).@def)
mod (palace(icl>building).@def,
royal(mod<thing))
plc (palace(icl>building).@def,
Madrid(iof>city))
[/S]
  
```

This is an explicit representation of the sentence meaning, in which an action (carried out by an agent (agt)) is described as one of *managing*. Such action is performed by an institution named ‘Turespaña’ and its object (obj) is a *restoration* activity. It is also specified that the object (obj) of that restoration is a *palace*, a type of building, located in Madrid. This representation contains a very precise characterization of the semantic relations and of the nature of the concepts present in the sentence. Such characterization is readily exploitable by an intelligent question answering system. This approach has been put into practice in the AgroExplorer system.

4.2 AgroExplorer

AgroExplorer [21] is a language independent search engine with multilingual information access facility. AgroExplorer is a system of Information Retrieval that searches directly into UNL documents. Queries are also transformed into UNL, so that the search engine only deals with UNL representations and it is therefore completely language independent. AgroExplorer also exploits the multilingual capabilities of UNL when rendering queries results. As long as language generators are available, answers can be generated into different natural languages.

A complete description of this system can be found in [15] where the architecture and its modules are described. An on-line demo is also available at <http://agro.mlasia.iitb.ac.in/>

5 Conclusions

Apart from multilingual generation applications, UNL is currently being employed as document representation formalism in tasks such as information extraction [15], text summarization [16],

and integration with other linguistic ontologies [12], [17]. UNL should not be seen either as just another interlingua neither as just another knowledge representation formalism. Its goal is to serve as an intermediate *knowledge* representation that can be exploited by different knowledge intensive tasks.

UNL is a formalism worth to be considered particularly in those scenarios where:

1. Multilingual acquisition and dissemination of textual information is required,
2. Deep text understanding is required for providing advanced services such as question answering, summarization, knowledge management, knowledge-based decision support, language independent document repositories, etc. For all these tasks, a domain and task dependent knowledge base is needed and building it from UNL representations presents distinct advantages over other approaches.

References

- [1] C. Fellbaum, Ed, "WordNet: An Electronic Lexical Database", in *Language, Speech, and Communication Series*, MIT Press, 1998.
- [2] R. Sharma, and S. Raman: "A Phrase-Based Text Representation Approach for Effective Retrieval of Web Documents", in *Proceedings of the International Conference on Artificial Intelligence, IC-AI '03*, Las Vegas, Nevada, USA, 2003.
- [3] M. Holub, and A. Böhmová, "Use of Dependency Tree Structures for the Microcontext Extraction", in *Proceedings of ACL2000, Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, Hong Kong, 2000
- [4] D. Manjula, G. Aghila, and T. V. Geetha. "Document Knowledge Representation using Description Logics for Information Extraction and Querying", in *Proceedings of the International Conference on Artificial Intelligence, IC-AI '03*, Las Vegas, Nevada, USA, 2003.
- [5] H. Uchida, "ATLAS-II: A machine translation system using conceptual structure as an Interlingua", in *Proceedings of the Second Machine Translation Summit*, Tokyo, 1989.
- [6] K. Muraki, "PIVOT: Two-phase machine translation system". In *Proceedings of the Second Machine Translation Summit*, Tokyo, 1989.
- [7] E. H. Nyberg, and T. Mitamura, "The KANT system: fast, accurate, high-quality translation in practical domains", in *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, vol. 4, pp. 1254-1258, Nantes, 1992.
- [8] S. Beale, S. Nirenburg S. and G. Mahesh, "Semantic Analysis in the Mikrokosmos Machine Translation Project", in *Proceedings of the Second Symposium on Natural Language Processing (SNLP-95)*. Bangkok, Thailand. 1995.
- [9] B. Dorr, "Machine Translation Divergences: A Formal Description and Proposed Solution", *Computational Linguistics*, vol 20(4), pp 597-633, 1994.
- [10] Jackendoff, R., *Semantic Structures*. Current Studies in Linguistics series. Cambridge, Massachusetts: The MIT Press, 1990
- [11] H. Uchida, *The Universal Networking Language Specifications, v3.3*, 2004. Available at <http://www.undl.org>.
- [12] L. Iraola, "Using Wordnet for linking UWs to the UNL UW System", in *Research on Computing Science. Special Issue on UNL*, vol 12, pp 369-378, 2005.
- [13] I. Boguslavsky, J. Cardeñosa, C. Gallardo, L. Iraola. "The UNL Initiative: An Overview", *Lecture Notes in Computer Science*, vol 3406, pp 377 – 387, 2005
- [14] J. Cardeñosa, C. Gallardo, and L. Iraola. "An XML-UNL Model for Knowledge-Based Annotation". *Research on Computing Science. Special Issue on UNL*, vol 12, pp 300-308, 2005.
- [15] P. Bhattacharyya, et. al, *Agro-Explorer: A meaning based multilingual search engine*, available at: http://www.projects.mlasia.iitb.ac.in/docs/agro_icdl.htm.
- [16] V. Sornlertlamvanish, T. Potipiti. and T. Charoenporn, "UNL Document Summarization", in *Proceedings of the First International Workshop on Multimedia Annotation (MMA 2001)*, Tokyo, 2001.
- [17] C. Ribeiro, R. Santos, R. P. Chaves and P. Marrafa. "Semi-automatic UNL Dictionary Generation using Wordnet", In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 279-282, Lisbon, Portugal, 2004.