Preface

This volume[1] is an attempt to compile and illustrate all the open lines of research within the UNL initiative. The included papers constitute a selection of the most significant papers presented in several international conferences and workshops during the last four years that served as a meeting point for the UNL consortium. In general, papers are not restricted to UNL although they are clearly predominant; they clearly illustrate the wideness and flexibility of this UNL initiative, launched by the United Nations aiming at the elimination of linguistic barriers.

Since the starting of the UNL project in 1996, the participants in the project from initially 15 languages have made substantial progress in technical matters and the organizational aspects involved as well. This book attempts to provide a survey on the approaches and theoretical studies around UNL, since research on UNL is not only devoted to studies on interlinguas, MT or any NLP related issues, the intrinsic properties of UNL make it a firm candidate to support a wide variety of applications ranging from e-learning platforms to management of multilingual document bases. Such a variety of applications, their theoretical basis and subsequent methodological inquiries are at core of this volume.

What is UNL? Its motivation and purpose

The emerging needs and use of Internet for cultural and educational dissemination and commercial expansion of the peoples collide with linguistic diversity, which in principle diminishes the potential of Internet as a vehicle of knowledge for everybody. Aware of this problem, the Institute of Advanced Studies of the University of the United Nations University (UNU/IAS) launched the UNL project in 1996 with the initial participation of 15 languages (German, Arab, Chinese, Spanish, French, Hindi, Indonesian, English, Italian, Japanese, Latvian, Mongol, Portuguese, Russian, Thai). In short, the UNL Programme was initially conceived to support multilingual services in Internet being an alternative to classical machine translation systems.

The UNL system revolves around a unique artificial language (Universal Networking Language) that pretends to capture the meaning of written documents. This language is based on the representation of concepts and its relations. The definition of this language has been possible thanks to the collaboration of more than one hundred people, prestigious researchers, and scientists of all around the world, that worked during the first three years of the project to produce a final version of the UNL specifications[2].

[1] Earlier versions of the papers at pages 10, 109, 117, 125, 145, 215, 230, 254, 268, 276, 309, 347, 359, 370, 380 have been published in the Proceedings of Convergences'03, Alexandria, Egypt. Earlier versions of the papers at pages 3, 10, 27, 38, 101, 261, 326 have been published in the Proceedings of LREC-2002.

[2] UNL Specifications, v.3.1 available at
http://www.undl.org/unlsys/unl/UNL%20Specifications.htm

The UNL organization

The UNL initiative has often been regarded as “hidden organization”. The first years of the project (1996-2000) were devoted to the definition of the interlingua and to the development of the essential components required to undertake the basic process in UNL (mainly dictionaries and language generators). During this period, the organization was closed and limited to a number of participants, because of the need to define the specifications of the language.

By the end of this period, the UNL project reached a significant degree of quality in the development of components, linguistic resources and technical specifications; and the specifications were finally produced. Once the specifications were finished, they were made public and accessible to all the international community, so that collaboration and *participation in this initiative is completely open.*

As a consequence of this degree of development, the Board of the United Nations University, in its fifth meeting in 2000, agreed on the creation of a new institution responsible for the organization and promotion of the UNL in the future under the umbrellas of the United Nations. This new entity was the UNDL Foundation, with headquarters in Geneva[3]. The development of the components of different languages was assigned to the so-called Language Centres, constituted by the initial teams in each country in charge of the development of the essential components of UNL.

The year 2004 represents a turning point in the evolution of UNL for two main reasons. First, it is the year where a new period coordinated and fostered by the Language Centres starts for the debugging, updating and expansion of linguistic resources and developed components of their representative languages, in order to respond to the institutional and marketable challenges at a pre-competitive level in the support of multilingual services. Second, it is the year where the UNL patent has been approved in USA for the UN (US Patent No. 6704700 B1, March 2004). It has been the first software patent of the United Nations.

Open nature and scientific dissemination of UNL

Since 2002, an open annual conference around the convergence of language, culture and knowledge is being held as a meeting point for researchers, politicians, linguists and engineers. The most recent edition of this open conference was Convergences’03, held in Alexandria, Egypt. The most significant papers from this conference have selected and included in this volume. Additionally, an international workshop on UNL and Interlinguas was organized in 2002 (International Workshop on UNL, other Interlinguas and their Applications, held at Las Palmas de Gran Canaria, May, 2002), papers from this workshop are also compiled in this volume. Finally, we include the papers of the current edition of the UNL Workshop, held in Mexico D.F, February, 2005.

These conferences and workshops try to be a forum where all the interested people in this initiative find a vehicle for communication and exchange of knowledge. The

[3] www.undl.org

UNL is a great initiative that could never succeed and advance if the number of participants is limited to the initial ones. The heterogeneity of the authors and languages involved in this selection of papers shows the open nature of UNL.

Research on UNL: Current Trends

Apart from the mere applied studies of UNL, there is a current important trend on theoretical studies of UNL, even though there is a final version of the specifications of the language, dating to July 2003.

The rationale for such theoretical research is the need for standardization and homogenization on the use of the Interlingua both at the applied level and at the theoretical level. The UNL Specifications turned out to be subject to different personal interpretations, thus *creating own UNL dialects.* This is not desirable for an interlingua, that claims to be language independent and that, in fact, turned out to be "person-dependent". For this reason, it is important and desirable to foment theoretical studies on UNL, both from the linguistic point of view and the knowledge point of view.

From a scientific point of view, UNL follows the approach of the concept of Interlingua, as an "artificial" language aiming at the neutral representation of linguistic meaning. In this sense its roots can be sought in the tradition of MT interlinguas and in the tradition of Knowledge Representation formalisms.

When viewed as an interlingua, UNL differs from some of its predecessors and current Interlinguas in the generality of appliance, that is, UNL is not restricted to a number of languages or to a given domain. Thus, its design pretended to show the highest degree of language independence while retaining natural language expressiveness in order to support multilingual generation tasks.

Of course, the staging of UNL is such a general enterprise that requires research and efforts. This process can be divided into several periods:

- Creation of deconversion and enconversion modules, (see Part 3) that is, development of the basic tools to undertake the basic architecture of the UNL system (enconversion and generation), along with dictionaries. Although basic, it is *conditio sine qua non* to have powerful generation systems. This a fruitful trend in the UNL consortium, with three different approaches:

  1. The official one: those using a common engine provided by the UNL Center.
  2. The integrative ones: those that have integrated UNL into pre-existing MT systems, following the transfer-based architecture, showing the flexilibility of UNL with good results.
  3. The new ones: those that have noticed the drawbacks of the *official* components, and have decided to create new architectures for generation

  It should be noticed that emphasis is put on the deconversion process, quantitatively proven by the number of papers devoted to generation. Teams usually develop generation systems, not so much enconversion systems, although the integrative usually includes both processes in UNL.

- Application of UNL in other contexts (see Part 3). Should UNL be considered as an interlingua, it can be applied in fields and tasks other than multilingual generation, being the main one Knowledge representation and Knowledge Management.
- Use of external lexical and ontological resources. It is important as well, and following the spirit of the integrative approaches, the use of external lexical resources such as Wordnet to enhance some of the processes of UNL, especially in the lexicographic part (see Part 3, also). This is also a trend and the philosophy of UNL: integration and complementation of resources is encouraged, rather than confrontation. And this is the spirit of the consortium and of every work in UNL.

From an engineering point of view, research is taken on:

- Creation of methodologies in the workflow.
- Standardization of UNL, integration of UNL into current standards.

Why such studies methodologies and standards? Because of the heterogeneity and diversity of the current consortium, it is needed such a process of standardization and methodologies, since the short and medium term objective of UNL is its staging in the market, where standards and methodologies are required in order to pursue higher productivity and quality. The areas of linguistic engineering together with knowledge engineering are claiming for such methodologies and processes of standardization.

The Future

After some time developing components and systems to support the multilingual services, UNL researchers and new teams have discovered that the UNL could be support of other applications as crosslingual information retrieval, knowledge repositories, automatic building of ontologies from texts once repressented in UNL and much more. UNL could be useful in new possible applications in areas where a common conceptual representation is needed, independent of any particular language. For doing it, new necessities emerge; particularly when putting together semantics and multilingualism. More theoretical studies are needed, along with the tuning up of resources and tools, the proper standardization of the interlingua and processes for enconverting and deconverting, and of course the integration and definition of the lexical component of UNL.

The Structure of the Book

The volume is divided into four parts.

Part 1. Introduction

This fist part is an introduction to the language itself, and its purpose is to set up the reader in the UNL context. These introductory papers posit the general philosophy of the language (paper at page 3) and provide a general introduction to the language itself and to the context of multilingual generation, one of the main and most basic "applications" supported by UNL (paper at page 10).

Part 2. Fundamentals

This part is dedicated to theoretical studies on UNL. As already said, UNL is mainly an interlingua. There are many aspects that have to be taken into account when designing an interlingua, such as its expressiveness, degree of language-independency, accuracy and formality of the language, etc. Most of these issues are covered in this part. Thus, the part opens up with an experiment on the common understandability of UNL by different humans and the admissible degree of indeterminacy and ambiguity in an Interlingua (paper at page 27). Pure theoretical studies on the universality of UNL and its adequacy from a representational and linguistic point of view follow (papers at pages 51 to 101). It has to be pointed out that this part is not exclusively devoted to UNL, but to the field of interlinguas in general (paper at page 38; paper at page 109).

All these papers point at the proper designs of the Interlingua. However, there is another important aspect worth of consideration in any artificial language, namely, the syntactic formalism of the formal language and its adequacy to the declared purpose. These topics are addressed in papers at page 117 and at page 125, where the emphasis is put on the syntactic properties of UNL expressions and its consequences to other issues such as analysis or proper deconversion. Finally, there is a (recurrent) thematic shift; UNL is not viewed as an interlingua to support linguistic tasks, but as a language for knowledge representation (papers at page 138 and at page 145).

These two sides of UNL (an interlingua to support linguistic tasks and a as knowledge representation language) determine the nature of the applications dealt with in Part 3.

Part 3. Applications

The core applications of UNL are those that support the tasks of NL analysis and generation (*enconversion* and *deconversion* in the UNL jargon). When dealing with NLP tasks, the scene is quite heterogeneous: from the use of common generation tools provided by the UNL Center (as shown in papers at pages 215 and 241), to the integration of existing MT translation systems based on the transfer architecture to support an Interlingua architecture (papers at pages 157 and 230). Other languages are supported with new tools, but differs in their configuration and architecture (maybe reflecting language variety, maybe reflecting different ways to support generation and of course, as an advanced over common tools, like Deco). Chinese, Brazilian Portuguese, Arabic or Armenia are example of this, where very different paradigms are illustrated in order to undertake the generation task (papers at pages 167, 175, 195, and 210, respectively).

Papers at pages 254 to 276 illustrate the development of workbenches to support the processes of edition, generation and training and with the creation of multilingual platforms within the UNL framework.

In parallel with the theoretical studies of Part 2, UNL also presents and applied dimension when conceived as a language for knowledge representation (papers at pages 337 and 359). These papers present the use of UNL as an extension (or complementation) to the expressiveness of standard languages such as XML (illustrated in papers at pages 300 and 309), as the communication language among agents, developed in paper at page 326, or as the support of case-based reasoning systems (paper at page 347). It is also remarkable the possibility of complementation and integration with other lexical and ontological resources such as WordNet (papers at pages 370 and 380) to the enhancement of the processes of knowledge acquisition and representation within the UNL context. Finally, paper at page 286 shows how to extend the expressivity of UNL in order to represent and formalize meaning coming for oral sources.

Part 4. Methodologies

Finally, the volume ends up with the methodological work. Methodologies target at the creation of methodologies to support multilingual services (papers at pages 395 and 413) and for the optimization of knowledge intensive tasks (paper at page 430). Needless to say, methodologies conforms an integral part of the UNL R+D activities, as long as productivity, quality and a real consolidation of UNL are pursued both at the scientific and commercial levels.

Mexico D.F, 16th February 2005

Jesús Cardeñosa
Alexander Gelbukh
Edmundo Tovar
Editors

Table of Contents

Índice

INTRODUCTION: Setting up UNL

FOUNDATIONS

APPLICATIONS

METHODOLOGIES

Prologue

UNL is an ongoing worldwide initiative starting in 1996. Almost 10 years have passed a big span of time for a project. We could say that UNL didn't meet its expectations. But let's have a closer look to UNL, the project, its basics and objectives. A closer look at its objective will reveal that this affirmation is gratuitous and unmotivated.

The Problem: Linguistic Diversity

UNL was launched by IAS/UNU to erase linguistic barriers. Linguistic barriers collide with the enhancement of linguistic diversity and the value that native languages as one of the main vehicles to express one's cultural identity. Apart from socio-cultural issues, linguistic diversity also knows an economic and political dimension. Institutions like the United Nations or the European Union have to face everyday with the barriers that linguistic diversity imposes. It is well known the enormous amount of documentation that these institutions produce everyday, which have to be produced in all their official languages: 6 for the UN, 25 for the European Union. It is simply unfeasible to rely on human translators for the production of all these amount of documentation.

Aware of this, the IAS/UNU launched the UNL project, aiming at the real access of information in the own native language and not recurrent to dominant languages. UNL is basically an artificial language where contents expressed in natural languages can be converted to and subsequently, contents written in UNL can be generated into any natural language, provided that the adequate tools are built.

MT and Multilinguality

From the technological point of view, multilinguality has been tackled by Machine Translation. In the evolution of the area of MT, there is variety of architectures to undertake the task of translating the *contents* of one text written in a given language into another language. Transfer-based systems could be regarded as the most productive and of better quality. But they are hindered by the exponential growth in the modules to be developed when the number of involved languages increases. A transfer-based system involving N languages need to develop N*(N-1) modules. An astronomic number to create real multilingual platforms.

Further, although there are some very good systems, the quality of these systems seem to be limited, since after years of refinement, the MT system does not surpass a given degree of quality. Besides, the development of transfer based MT systems is usually reduced to the so-called majority languages (English, French, German and even Spanish or Italian), but it is fairly rare to find a good quality and wide coverage MT system covering English and Polish, let's say.

Transfer based MT is not the only option, Interlingua-based systems represents an alternative to transfer systems. Interlingua-based MT does not work on pair of languages, but translation is carries out to and from an artificial language that serves as a pivot for all the natural languages involved in the system. This architecture tries to overcome the exponential growth of transfer-based systems, since the number of modules to develop for N languages is 2*N and the inclusion of new languages into the system does not affect the other language modules. In this way, UNL follows the architecture of Interlingua-based MT systems.

Usually, Interlinguas are abstract formal (or semi formal) languages that captures the meaning of texts in a language independent way. Ideally, the Interlingua should not be close to a given particular language and should not include linguistic devices proper of natural languages. In this way, Interlingua-based systems seem the most plausible (and even the unique) option to tackle massive multilinguality.

But Interlinguas has been often rejected within the scientific community and since their boom in the 80ies, there have no commercial application of Interlinguas and the systems developed under this trend were laboratory products. Why is this so? Let's have a look at the properties of interlinguas.

Problems with Interlinguas

Interlinguas are *semantic languages* designed to represent the meaning of any given text, ideally satisfying the following conditions:

(a) They are language neutral.
(b) They are precise, unambiguous, formal languages

Being so, they usually show the following characteristics:

- Interlinguas are intimately tied up with ideas about the representation of meaning, being meaning the most abstract and deepest level of linguistic analysis (that should be common to all languages, far enough from surface representation of languages).
- An Interlingua is "another language" in the sense that it has autonomy and thus its components need to be defined: vocabulary and "relations" mainly. Besides, and Interlingua is an artificial language that should be as expressive as natural languages.

Here we find the main bottleneck of interlinguas: its proper design and definition. Defining an Interlingua involves the following parameters:

(a) A language whose "atoms" are not dependent on any given natural language so that the ambiguity of natural languages is eliminated.
(b) A language whose "atoms" are not dependent on a given natural language so that the concepts and ideas expressed in different natural languages can be easily and naturally expressed in the Interlingua.
(c) A language that is as expressive as a natural language so that what can be expressed in natural languages can be transposed to the Interlingua, and from the interlingua to other natural languages.

These three conditions make interlinguas hard to design. It is quite difficult to find the equilibrium between language independency, degree of abstraction and expressiveness in a formal device such an Interlingua. Maybe this difficulty in the design of interlinguas is the reason why they have not been successful at least in open domains within massively multilingual environments. The examples of interlingua-based systems are domain dependent and quite limited in the number of languages.

Is UNL a Viable Solution?

The panorama appears quite despairing. While Interlinguas are theoretically biased and difficult to put into practice, transfer based systems have proved to be unattainable when dealing with massive multilinguality. Maybe the concept of Interlingua should be revisited, and re-adapted to real necessities and to real scenarios. This is the spirit of UNL. UNL, by its definition and by its most basic architecture is definitely an Interlingua-based system. Its targets are the support of multilinguality, not restricted to a given domain or to a given family of languages. Thus, the design of a interlingua like UNL encounters all the possible barriers that an Interlingua may encounter (especially to find a real language independent representation).

So why we could considered UNL as different, as a new viable technology if interlinguas were rejected a long time ago? First, let's remember the main objective of UNL:

- to generate and produce contents in any natural language in any domain.
- to support multilingual services.

That is, there is a primacy of generation and coverage of languages and domains, which means that a **very expressive formalism** has to be designed in order to represent such a variety of contents coming from any natural language.

Let's illustrate this fact by have a closer look at the vocabulary of the Interlingua, one of the most difficult and polemic issues of UNL and of any Interlingua. UNL utilizes the so-called Universal Words as the semantic atoms of the Interlingua (no decomposable). They exhibit the following main characteristic:

They are based on English headwords.

From this very simple definition, we can conclude that UNL is language biased (English) and thus:

1. UNL is based on a natural language:
2. It hinders logical relations and inferences (facilitated by primitive based solutions)
3. Its vocabulary is a potential source of ambiguity
4. Its vocabulary fosters lexical and conceptual mismatches among languages.

So is there any advantage in the UW system and in the overall essence of UNL? Well, if theoretical reasons do not support the design of open-domain interlinguas, <u>let's look at the practical or pragmatic ones.</u>

(a) UNL is based on a natural language. At first sight could be a drawback, however, the expressiveness of a natural language is inherited by the Interlingua, thus allowing for the representation of a variety of domains and concepts.

(b) UNL shows an English oriented vocabulary. At this moment, English is the lingua franca, the most accessible to work with for Indo-Europeans, Semitic, Japanese, Chinese, etc. Bilingual dictionaries usually have English as one of their target/source languages, thus the development of lexicographic resources is facilitated by choosing English as the most basic atoms of the language.

Of course, this approach (although supported by pragmatism) is far from perfect. Even at first sight, it can be considered as naïve, since it merely "suggest" well known problems in lexical semantics (like support verbs, compounds expressions, connotational meaning, etc). For this reason, theoretical research on the UNL as a language itself should be fostered within the Consortium, while respecting the basic nature of the language.

That is, UNL should be viewed rather than a perfect Interlingua as **the pillars to support multilingual services**. Its natural language orientation (apparently, its weakest points as an Interlingua) turns the language as a candidate to the support of multilinguality and facilitates converting contents to and from UNL. There are several aspects that support it. First, the creation of generators of medium quality (where post-edition is possible) is rather straightforward. Second, its flexibility and language orientation makes it possible to integrate UNL into other pre-existent MT systems (be it transfer-based be it another architecture) which extends the range of application of UNL and makes possible to alleviate the problem of exponential growth in transfer-based systems. And last, but not least, the processes of enconverting and deconverting are independent so that if generation is taken as a priority, generators are constructed first; the process of enconversion can be done manually, due to the human readability of the language.

At this point in the evolution of UNL, there appears a contradiction, UNL is still not theoretically mature, but from an applied perspective, it is. In the short term there is priority for the UNL Consortium to get feedback from previous experiences in Interlinguas, from Linguistic Theory (semantics, logic, and lexical semantics) in order for UNL to grow and find a place in the scientific community and, why not, in the market as a real approach to support multilinguality, once the applications and utilities are clear and defined within the UNL Programme.

Prospective

So is it worth another attempt? Definitely yes, the real need to overcome linguistic barriers (be it at the institutional level, be it at the social level) claims for a solution to the problem of multilinguality. Transfer based systems simply are out of question *if isolated*. This doesn't mean that they are useless: they are not. An interlingua like UNL is conceived as another autonomous languages, close enough to the superficial form of natural languages, thus integration of the Interlingua into the transfer system is possible and not a *contradiction in terminis*.

After several years of experience, we know that knowledge and language generation do not go *on a par*. Thus the final design have to be done bearing the ultimate purpose of the interlingua (the closer to language semantics is, the better to generate languages) and probably will lead to the success of the interlingua.

A Final Word

I would like to thank the editors of this book for their invitation to write a prologue to this work and to collaborate with them in the selection and revision of the selected papers presented in this volume. Hopefully it will provide a thorough understanding of the UNL Programme, its meaning, its evolution, its shortages and its strengths.

Carolina Gallardo

A Rationale for Using UNL as an Interlingua and More in Various Domains

Christian Boitet

GETA, CLIPS, IMAG
385, Av. de la Bibliothèque, BP 53
F-38041 Grenoble cedex 9, France
Christian.Boitet@imag.fr

Abstract. The UNL *language* of semantic graphs may be called as a "semantico-linguistic" interlingua. As a successor of the technically and commercially successful ATLAS-II and PIVOT interlinguas, its potential to support various kinds of text MT is certain, even if some improvements would be welcome, as always. It is also a strong candidate to be used in spoken dialogue translation systems when the utterances to be handled are not only task-oriented and of limited variety, but become more free and truly spontaneous. Finally, although it is not a true representation language such as KRL and its frame-based and logic-based successors, and although its associated "knowledge base" is not a true ontology, but rather a kind of immense thesaurus of (interlingual) sets of word senses, it seems particularly well suited to the processing of multilingual information in natural language (information retrieval, abstracting, gisting, etc.).The UNL *format* of multilingual documents aligned at the level of utterances is currenly embedded in html (call it UNL-html), and used by various tools such as the UNL viewer. By using a simple transformation, one obtains the UNL-xml format, and profit from all tools currently developed around XML. In this context, UNL may find another application in the localization of multilingual textual resources of software packages (messages, menu items, help files, and examples of use in multilingual dictionaries.)

1 Introduction

UNL is the name of a project, of a meaning representation language, and of a format for "perfectly aligned" multilingual documents. There is some hefty controversy about the use of the UNL language as an "interlingua", be it for translation or for other applications such as cross-lingual information retrieval. On the other hand, there is almost no discussion on the UNL format, in its current form, embedded in HTML, or some directly derivable form, embedded in XML.

We argue that the UNL language is indeed a good interlingua for automated translation, ranging from fully automatic MT to interactive MT of several kinds through, we believe, spoken translation of non task-oriented dialogues. It is also more than that, due to the associated "knowledge base", and has a great potential in textual information processing applications.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 3–9.

Standardization of the Generation Process in a Multilingual Environment

Jesús Cardeñosa, Carolina Gallardo and Edmundo Tovar

Universidad Politécnica de Madrid, 28660 Madrid.
{carde,carolina,edmundo}@opera.dia.fi.upm.es

Abstract. Natural language generation has received less attention within the field of Natural language processing than natural language understanding. One possible reason for this could be the lack of standardization of the inputs to generation systems. This fact makes the systematic planning of the process of developing generation systems to become difficult. The authors propose the use of the UNL (Universal Networking Language) as a possible standard for the normalization of inputs to generation processes.

1 Introduction

In natural language processing (from now on NLP) two areas can be differentiated: analysis and generation. However, one has not received the same attention as the other from the scientific community, that is why generation can be considered as the "poor brother" of the NLP. The reason for this minor development is the different nature of the input to the analysis and generation systems. The input to the analysis systems is always natural language, whose casuistic and phenomenology are known; while in a generation system, the output is always known, but not what it is going to generate from [1].

The input to a generation system varies depending on whether it is monolingual generation (dialogue systems) or a multilingual system (mainly machine translation systems). In dialogue systems it is difficult to establish appropriate characteristics common to all inputs, because "the problem" of generation is usually solved with solutions *ad hoc*, depending on the application and the system language. In machine translation systems, there are also many differences in the inputs to the generation subcomponents, conditioned by the nature of system architecture (transfer, interlingua, etc.), the kind of grammars being used (declaratives vs. procedural) [2], or the number of languages in the system.

This difference in the input to the generators makes a systematic planning of their development process impossible (main cause of the minor development of generation compared to analysis). It is necessary then, that the input to the "generator" can be supported with an appropriate model of contents representation, separated from the format or language that ensures a standard process for the development of generation systems.

In this article we propose the UNL as a possible standard for the generation inputs. To achieve this, in section 2 we will introduce the main generation architectures. Sec-

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 10–24.

The UNL Distinctive Features: Inferences from a NL-UNL Enconverting Task

Ronaldo Teixeira Martins,[1] Lúcia Helena Machado Rino,[2]
Maria das Graças Volpe Nunes,[3] Osvaldo Novais Oliveira Jr.[4]

[1]Núcleo Interinstitucional de Lingüística Computacional - NILC
Av. do Trabalhador São-Carlense, 400 - 13560-970 - São Carlos, SP, Brazil
ronaldo@nilc.icmc.sc.usp.br

[2]Departamento de Computação - Centro de Ciências Exatas e de Tecnologia - UFSCar
Rod. Washington Luiz, km 235 - Monjolinho - 13565-905 - São Carlos, SP, Brazil
lucia@dc.ufscar.br

[3]Instituto de Ciências Matemáticas e da Computação (ICMC) - Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400 - 13560-970 - São Carlos, SP, Brazil
mdgvnune@icmc.sc.usp.br

[4]Instituto de Física de São Carlos (IFSC) - Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400 - 13560-970 - São Carlos, SP, Brazil
chu@ifsc.sc.usp.br

Abstract. This paper reports on the distinctive features of the Universal Networking Language (UNL). We claim that although UNL expressions are supposed to be unambiguous, UNL itself is able to convey vagueness and indeterminacy, as it allows for flexibility in enconverting. The use of UNL as a pivot language in interlingua-based MT systems is also addressed.

1 Introduction

Machine Translation (MT) is one of the most controversial subjects in the field of natural language processing. Researchers and developers are often at odds on issues concerning MT systems approaches, methods, strategies, scope, and their potentialities. Dissent has not hindered, however, the establishment of tacit protocols and core beliefs in the area. It has often been claimed that:[1] (1) fully automatic high-quality translation of arbitrary texts is not a realistic goal for the near future; (2) the need of some human intervention in pre-edition of the input text or in post-edition of the output text is mandatory; (3) source language should be rather a sublanguage, and the input text should be domain- and genre-bounded, so that the MT system could cope with natural language ambiguity; (4) the transfer approach is more feasible than the interlingual one, since the latter, albeit more robust and economic, is committed to the

[1] Most of these assumptions can be extracted from the Survey on the State of the Art in Human Language Technology (Cole et al., 1995). Of special interest are the articles concerning multilinguality by Martin Kay (8.1, 8.2) and Christian Boitet (8.3, 8.4).

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 27–37.

Issues in Generating Text from Interlingua Representations

Stephan Busemann

DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
busemann@dfki.de

Abstract. Multi-lingual generation starts from non-linguistic content representations for generating texts in different languages that are equivalent in meaning. In contrast, cross-lingual generation is based on a language-neutral content representation which is the result of a linguistic analysis process. Non-linguistic representations do not reflect the structure of the text. Quite differently, language-neutral representations express functor-argument relationships and other semantic properties found by the underlying analysis process. These differences imply diverse generation tasks. In this contribution, we relate multi-lingual to cross-lingual generation and discuss emergent problems for the definition of an interlingua.

1 Introduction

In this contribution, we relate multi-lingual to cross-lingual generation and discuss emerging problems for the definition of an interlingua. Multi-lingual generation starts from non-linguistic content representations for generating texts in different languages that are equivalent in meaning. The generation of weather forecasts or environmental reports are typical examples. In contrast, cross-lingual generation is based on a language-neutral content representation which is the result of a linguistic analysis process. Generation for machine translation is a most prominent example.

Non-linguistic representations do not specify linguistic semantics nor do they reflect the structure of the text to be generated. In contrast, language-neutral representations express functor-argument relationships and other semantic properties found by the underlying analysis process. These differences imply diverse generation tasks.

However, there are also commonalities. In both cases, generation is the mapping of some semantic representation onto linguistic strings. We may assume a single generation process that uses different separately defined language specific knowledge sources. In both cases, we may view the underlying representation as an interlingua, since it attempts to cross the language barrier by providing content descriptions independently of the target language.

An instance of each type of tasks has been implemented using the generation system TG/2 (Busemann, 1996), quickly overviewed in Section 2. The usage of the same framework allows us to relate the tasks to each other (Section 3) and to gain insights

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 38–50.

On the Aboutness of UNL

Ronaldo Teixeira Martins,[1,2] Maria das Graças Volpe Nunes.[2,3]

[1]Faculdade de Filosofia, Letras e Educação – Universidade Presbiteriana Mackenzie
Rua da Consolação, 930 – 01302-907 – São Paulo – SP – Brazil

[2]Núcleo Interinstitucional de Lingüística Computacional (NILC)
Av. Trabalhador São-Carlense, 400 – 13560-970 – São Carlos – SP – Brazil

[3]Instituto de Ciências Matemáticas e da Computação – Universidade de São Paulo
Av. Trabalhador São-Carlense, 400 – 13560-970 – São Carlos – SP – Brazil

ronaldomartins@mackenzie.com.br, gracan@icmc.usp.br

Abstract. This paper addresses the current status, the structure and role of the UNL Knowledge Base (UNLKB) in the UNL System. It is claimed that the UNLKB, understood as the repository where Universal Words (UWs) are named and defined, demands a thorough revision, in order to accomplish the self-consistency requirement of the Universal Networking Language (UNL). In order to emulate human cognition and constitute the "aboutness" of the UNL, the UNLKB should be decentralized, distributed and reorganized as a network of networks, allowing for multicultural information and dynamic data.

1 Introduction

The Universal Networking Language (UNL) is an "electronic language for computers to express and exchange every kind of information" [Uchida, Zhu & Della Senta, 1999]. It can be defined as a knowledge-representation formalism expected to figure either as a pivot language in multilingual machine translation (MT) systems or as a representation scheme in information retrieval (IR) applications. It has been developed since 1996, first by the Institute of Advanced Studies of the United Nations University, in Tokyo, Japan, and more recently by the UNDL Foundation, in Geneva, Switzerland, along with a large community of researchers—the so-called UNL Society—representing more than 15 different languages all over the world.

Formally, the UNL is a semantic network believed to be logically precise, humanly readable and computationally tractable. In the UNL approach, information conveyed by natural language utterances is represented, sentence by sentence, as a hyper-graph composed of a set of directed binary labeled links (referred to as "relations") between nodes or hyper-nodes (the "Universal Words", or simply "UW"), which stand for concepts. UWs can also be annotated with attributes representing context-dependent information.

As a matter of example, the English sentence 'Peter kissed Mary?!' could be represented in UNL as (1) below:

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 51–63.

A Comparative Evaluation of UNL Participant Relations using a Five-Language Parallel Corpus

Brian Murphy and Carl Vogel

Brian.Murphy@cs.tcd.ie*, Vogel@cs.tcd.ie
Department of Computer Science, University of Dublin, Trinity College

Abstract. In this paper we describe a manual case study in interlingual translation among five languages. Taking the UN Declaration of Human Rights in Chinese, English, German, Irish and Spanish, we annotated the five texts with a common interlingual logical form. We then studied four inventories of semantic roles (developed for both theoretical andNLP applications), including a subset of UNL's relations, and evaluated their suitability to describe the predicate-argument relationships found in the annotation. As a result, we make some suggestions for possible additions to the UNL relations, and propose that some of the existing relations be conflated or redefined.

1 Introduction

The work described here is part of a feasibility study on the use of semantic roles in interlingua-based machine translation. Our objective was to see if any set of semantic roles could give a description of verb-predicate relationships across a range of languages that would form an adequate basis for automatic generation.

The languages chosen were those that the authors have some working knowledge of (English, Chinese, German, Irish and Spanish), and include widespread and minority languages, both well and less-studied. The corpus used is theUN Declaration of Human Rights [1], a short text covering a broad range of topics in many languages (see Sect. 2).

From the literature on roles we selected four inventories (of whichUNL's relations is one) that we considered to be well-enough developed for the annotation of unrestricted text. These inventories ([2,3,4,5] detailed in Sect. 4) were also chosen to be representative both theoretically and in terms of application to tasks such as machine translation and information retrieval.

After aligning the five language versions of the corpus, we manually annotated each article of the text with a language-neutral logical form (effectively a prototype interlingua) following the guidelines described in Sect. 3.1. The main part of the work then involved applying each of the role inventories in turn to the logical form and determining whether they satisfied three key criteria: coverage, differentiation and lack of ambiguity (Sect. 5). In other words, one should

* Supported by the TCD Senior Lecturer's Broad Curriculum Fellowship and Enterprise Ireland

Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 64–76.

Some Controversial Issues of UNL: Linguistic Aspects

Igor Boguslavsky

Universidad Politécnica de Madrid (Spain)/IITP RAS (Russia)
igor@opera.dia.fi.upm.es

Abstract. We discuss several linguistic aspects of the Universal Networking Language (UNL); in particular, those connected with Universal Words (UWs), UNL relations, and hypernodes. On the one hand, the language should be rich enough and provide sufficient means to express the knowledge that might be required in the applications it is intended for. On the other hand, it should be simple enough to allow uniform and consistent use across languages and by all encoders. The major expressive device of UNL used for overcoming lexical divergence between languages is so-called restrictions. They have three functions, which are relatively independent of each other: the ontological function, the semantic function, and the argument frame function. We discuss various types of restrictions and propose new expressive means for describing UWs. Sample dictionary entries are given which incorporate our proposals. We propose several new UNL relations and discuss when and how hypernodes should be introduced.

1 Background

Among many problems that developers and users of a meaning representation language are facing, two somewhat conflicting requirements are standing out. On the one hand, the language should be rich enough and provide sufficient means to express the knowledge that might be required in the applications it is intended for. The more complex and knowledge-demanding the application, the more complex the design of the meaning representation language becomes. On the other hand, it should be simple enough to allow uniform and consistent use across languages and by all encoders. In the case of UNL, the latter problem is particularly serious, since the encoders work in different countries, belong to different linguistic schools, and have different linguistic traditions. Therefore, uniform understanding and use of UNL by all partners is difficult to achieve.

Since the start of the project in 1996, a large number of UNL-encoded documents have been accumulated that were produced by the project participants from 16 language groups each working on its native language. The analysis of these documents clearly shows two things: *UNL is still lacking means to express meaning adequately*, and *there is not enough uniformity in the UNL use among the partners*. To some extent, UNL has developed its own dialects. Despite the existence of the UNL Specifications, divergences between the dialects tend to grow. This tendency clearly manifests itself in the fact that all deconverters (=generators) are doing much better when dealing with the UNL documents produced by the authors of the deconverter than

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 77–100.

Some Lexical Issues of UNL

Igor Boguslavsky

Institute for Information Transmission Problems, Russian Academy of Sciences
19, Bolshoj Karetnyj, 101447, Moscow, Russia
bogus@iitp.ru

Abstract. The Universal Networking Language (UNL) developed by Dr. H. Uchida at the Institute for Advanced Studies of the United Nations University is a meaning representation language designed for multi-lingual communication in electronic networks, information retrieval, summarization and other applications. We discuss several features of this language relevant for correct meaning representation and multi-lingual generation and make some proposals aiming at increasing its efficiency.

1 UNL Approach to the Lexicon

The Universal Networking Language (UNL) developed by Dr. H. Uchida at the Institute for Advanced Studies of the United Nations University is a meaning representation language designed for multi-lingual communication in electronic networks, information retrieval, summarization and other applications.

Formally, a UNL expression is an oriented hypergraph that corresponds to a natural language sentence in the amount of information conveyed. The arcs of the graph are interpreted as semantic relations of the types agent, object, time, reason, etc. The nodes of the graph can be simple or compound. Simple nodes are special units, the so-called Universal Words (UWs) which denote a concept or a set of concepts. A compound node (hypernode) consists of several simple or compound nodes connected by semantic relations.

In addition to propositional content ("who did what to whom"), UNL expressions are intended to capture pragmatic information such as focus, reference, speaker's attitudes and intentions, speech acts, and other types of information. This information is rendered by means of attributes attached to the nodes.

After 6 years of the UNL project development, it is possible to take stock of what has been achieved and what remains to be done. In this presentation, I am going to concentrate on one of the central problems with which any artificial language is faced if it is designed to represent meaning across different natural languages. It is a problem of the language vocabulary.

I would like to single out three distinctive features of the UNL dictionary organization.

1. **Flexibility.** There is no fixed set of semantic units. There is only a basic semantic vocabulary that serves as a building material for free construction of derivative

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 101–108.

The Representation of Complex Telic Predicates in Wordnets: the Case of Lexical-Conceptual Structure Deficitary Verbs

Palmira Marrafa

University of Lisbon, Faculdade de Letras de Lisboa – DLGR
Alameda da Universidade, 1600-214, Lisboa
Palmira.Marrafa@netcabo.pt

Abstract. This paper has a twofold aim: (i) to point out that telicity is both a lexical and a compositional semantic feature; (ii) to propose a straightforward solution to represent lexical telicity in wordnets-like computational lexica. The approach presented here subsumes the basic idea that lexicon is not a repository of idiosyncrasies. It is rather organized following a few general (universal or parametrical) constraints. In this context, despite the fact that the paper is mainly concerned with Portuguese, cross-linguistic generalizations can be captured, on the basis of a contrastive examination of data. The analysis focus on the behavior of complex telic predicates, in particular those which are deficitary with regard to their lexical-conceptual structure. In order to represent appropriately such predicates in wordnets, the specification of information regarding semantic restrictions, within the corresponding synsets, is proposed as well as a telic state relation.

1 Introduction

Telicity is mostly considered a compositional property of meaning. This paper attempts to make evident it is also a lexical feature and, as a consequence, it has to be represented in the lexicon. A concrete proposal to encode telic information of complex predicates in wordnets is provided.

This proposal emerges from the need of representing the predicates referred to in the Portuguese WordNet (WordNet.PT), which is being developed in the EuroWordNet framework.

From an empirical point of view, the work presented here mainly deals with complex telic predicates, in particular with those which involve lexical-conceptual structure (LCS) deficitary verbs, in the sense defined in previous work (cf. [4] and [5]).

The paper is divided in three main sections: the first one briefly describes the EuroWordNet model; the second one discusses the lexical-conceptual structure (in the sense of [7]) of complex predicates on the basis of a semantics of events, arguing for the lexical nature of telicity, and adduces evidences supporting the idea that some verbs define a deficitary lexical-conceptual structure; finally, the third main section presents an integrated proposal to encode LCS deficitary verbs and their troponyms in wordnets.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 109–116.

Remaining Issues that Could Prevent UNL to be Accepted as a Standard

Gilles Sérasset and Étienne Blanc

GETA-CLIPS, IMAG, Université Joseph Fourier,
BP 53, 38041 Grenoble cedex 9,
Gilles.Serasset@imag.fr

Abstract. This paper presents practical issues when dealing with UNL (Universal Networking Language) documents. Some of these issues are at a purelly syntactical level, others are at a semantic level. Some of these issues introduce unnecessary difficulties when developing tools to handle UNL documents when others introduce unnecessary difficulties when encoding natural language utterances into UNL graphs.

1 Introduction

After several years of development, UNL (Universal Networking Language, [1, 2]) has proved its viability as a cross lingual data exchange format. Its expressive power makes it very useful for the development of multilingual information systems where it serves as a way to represent utterances in a language free manner. However, in order to be adopted as a standard, the UNL definition should be clarified or corrected in order to avoid common errors and misunderstandings.

As a UNL partner since 1998, the GETA (Groupe d'Étude pour la Traduction Automatique) group of the CLIPS (Communication Langagière et Interaction Personne-Système) lab develops and maintains a UNL deconverter for French. For this development, we are one of the few groups that decided to use our own existing tools (namely the ARIANE-G5 translator generator, [3–6]). As such, we had to develop several tools to parse and handle UNL documents and went accross some of the problems that will arise when UNL will be used by third party developers.

This paper presents some of the issues we faced and suggests some solutions. Our goal is to give UNL the opportunity to be largely adopted by third parties as a de-facto standard. After briefly presenting the UNL language and an example of an UNL document, we will begin by low level problems posed by the UNL syntax. After that, we will focus on middle level aspects involved when interpreting the UNL language at its computational level. Finally, we will present some of the higher level issues arising when we interpret UNL utterances as linguistic structures.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 117–124.

Semantic Analysis through Ant Algorithms, Conceptual Vectors and Fuzzy UNL Graphs

Mathieu Lafourcade

LIRMM, Université Montpellier II, 161, rue Ada,
34392 Montpellier cedex 5
mathieu.lafourcade@lirmm.fr

Abstract. In the context on the UNL project, we focus on the automatization of enconversion process, that is the building of UNL graphs from sentences. We present an extension of the UNL graph structure aiming at handling lexical and relational ambiguities. On this intermediate structure, we can apply ant algorithm propagation of conceptual vectors and other constraints. Graph nodes and relations have a level of excitement and when this level remains too low for too long they are deleted. This way, both acception and attachment selections can be performed.

1 Introduction

In itself, a text constitutes a complex system, but the computational problem is that the meanings are not strictly speaking active elements. In order to ensure the dynamicity of such a system, an active framework made of "meaning transporters" must be supplied to the text. These "transporters" are intended to allow the interactions between text elements and they have to be both light (because of their possible large number) and independent (word meanings are intrinsic values). Moreover, when some meanings stemmed from different words are compatible (*engaged* with *job* for instance), the system has to keep a trace of this fact. These considerations led us to adopt ant algorithms. Ant algorithms or variants of them have been classically used for optimisation problems like traveling salesman problem [*Dorigo* et al. 1997] among many others, but they were never used in Natural Language Processing (most probably because the NLP community contrary to the psycho-linguistics one, considered semantic aspects not very often as an optimization problem, nor explicitely modeled then as a dynamic complex system, [*Kawamoto* 1993] being a notable exception). However, [*Hofstadter* 1995] with the COPYCAT project, presented an approach where the environment by itself contributed to solution computation and is modified by an agent population where roles and motivations vary. Some properties of these models seem to be adequate for the task of semantic analysis, where word senses can be seen as more or less cooperating. We retain here some aspects that we consider as being crucial: (1) mutual information or semantic proximity is one key factor for lexical activation, (2) the syntactic structure of the text can

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 125–137.

Term-Based Ontology Alignment

Virach Sornlertlamvanich, Canasai Kruengkrai, Shisanu Tongchim,
Prapass Srichaivattana, and Hitoshi Isahara

Thai Computational Linguistics Laboratory
National Institute of Information and Communications Technology
112 Paholyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand
{virach,canasai,shisanu,prapass}@tcllab.org, isahara@nict.go.jp

Abstract. This paper presents an efficient approach to automatically align concepts between two ontologies. We propose an iterative algorithm that performs finding the most appropriate target concept for a given source concept based on the similarity of shared terms. Experimental results on two lexical ontologies, the MMT semantic hierarchy and the EDR concept dictionary, are given to show the feasibility of the proposed algorithm.

1 Introduction

In this paper, we propose an efficient approach for finding alignments between two different ontologies. Specifically, we derive the source and the target ontologies from available language resources, i.e. the machine readable dictionaries (MDRs). In our context, we consider the ontological concepts as the groups of lexical entries having similar or related meanings organized on a semantic hierarchy. The resulting ontology alignment can be used as a semantic knowledge for constructing multilingual dictionaries.

Typically, bilingual dictionaries provide the relationship between their native language and English. One can extend these bilingual dictionaries to multilingual dictionaries by exploiting English as an intermediate source and associations between two concepts as semantic constraints.

Aligning concepts between two ontologies is often done by humans, which is an expensive and time-consuming process. This motivates us to find an automatic method to perform such task. However, the hierarchical structures of two ontologies are quite different. The structural inconsistency is a common problem [1]. Developing a practical algorithm that is able to deal with this problem is a challenging issue.

The rest of this paper is organized as follows: Section 2 discusses related work. Section 3 provides the description of the proposed algorithm. Section 4 presents experimental results and findings. Finally, Section 5 concludes our work.

2 Related Work

Chen and Fung [2] proposed an automatic technique to associate the English FrameNet lexical entries to the appropriate Chinese word senses. Each FrameNet

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 138–144.

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.

1 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 con-

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 145–154.

About and Around the French Enconverter and the French Deconverter

Etienne Blanc

GETA, CLIPS-IMAG
BP53, F-38041 Grenoble Cedex 09
etienne.blanc@imag.fr

Abstract. We briefly describe the French Enconverter and the French Deconverter. We discuss then a few general points concerning the possibility of designing dependency trees equivalent to UNL graphs, the treatment of the ambiguity and anaphora resolution, and the structure of the compound nodes.

1 Introduction

In a previous paper [1], we described the basic principle of our French Enconverter, in which the UNL input graph is processed into an equivalent Dependency Tree, which is in turn applied to the entry of a rule-based French generator. We developed similarly a French enconverter, in which a French Analyser provides a representation of the text meaning as a Dependency Tree, which is further processed into an equivalent UNL graph.

In this paper, we will first briefly present the structure of the French Deconverter and Enconverter. We will then recall and discuss a little further than in our previous paper the general problem of the equivalency between UNL graph and dependency tree. And finally briefly comment on three topics we had to deal with when devising our Enconverter and Deconverter : Ambiguity and Anaphora Resolution, Processing of the Unknown Word, the exact structure of the Compound Node of a UNL graph.

2 Overall Structure of the French Deconverter and Enconverter

The French Enconverter and the French Deconverter are written on ARIANE-G5.

ARIANE-G5 is a generator of MT systems, which is an integrated environment designed to facilitate the development of MT systems. These MT systems are written by a linguist using specialised languages for linguistic programming. ARIANE is not devoted to a particular linguistic theory. The only strong constraint is that the structure representing the unit of translation (sentence or paragraph) must be a decorated tree.

Fig.1 shows an overview of a classical transfer MT system using the ARIANE environment. The processing is performed through the three classical steps: analysis, transfer and generation. An interactive disambiguation module may be inserted after

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 157–166.

A UNL Deconverter for Chinese

Xiaodong Shi, Yidong Chen

Institute of Artificial Intelligence
School of Information Sciences and Technologies, Xiamen University
361005 Xiamen, China
{mandel, ydchen}@xmu.edu.cn

Abstract. This paper describes the internal working of a novel UNL converter for the Chinese language. Three steps are involved in generating Chinese from UNL: first, the UNL expression is converted to a graph; second, the graph is converted to a number of trees. Third, a top-down tree walking is performed to translate each subtree and the results are composed to form a complete sentence. Because each node is visited exactly once, the algorithm is of linear time complexity and thus much faster than the standard deconverter provided by the UNL center. A manual evaluation effort was carried out which confirmed that the quality of the Deconverter output was better than that of the standard deconverter.

1 Introduction

Although the UNL [1],[2] center provides a language independent generator [3] which can deconvert UNL expressions into any language provided that a UW dictionary, a set of deconversion rules, and optionally a co-occurrence dictionary are available for that language, that deconverter has a number of deficiencies: First, the deconversion rules are rather difficult to write because of the cryptic formats imposed by the deconversion specification. Second, although the power of the deconverter is claimed to be that of the Turing machine [4], its speed is rather slow and thus unsuitable for the main web application, embedded multilingual viewing of a UNL document that is one of the key goals of the UNL. Third, most importantly, the deconversion software is not open-sourced, so that fixing any bugs or introducing much-needed improvements is at the mercy of the UNL center, which has been rather lacking in technical support and in releasing new versions. So we think it is necessary to develop our own deconverter for Chinese. This paper describes such an endeavor. However, it should be noted that although we concentrate on generating Chinese from the UNL expressions, nothing in our deconverter is inherently related to Chinese, thus the deconverter is also language independent.

This paper is organized as follows: Section 2 will describe the main components of the deconverter and the algorithms involved. Section 3 will focus some issues in generation, especially those related to the Chinese language, and in Section 4 we will briefly discuss related work in the literature. In Section 5 we will give example uses of the deconverter and finally we will present the conclusions.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 167–174.

Flexibility, Configurability and Optimality in UNL Deconversion via Multiparadigm Programming

Jorge Marques Pelizzoni, Maria das Graças Volpe Nunes

Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação
Av. do Trabalhador São-Carlense, 400. CEP 13560-970. São Carlos – SP – Brasil
{jorgemp, gracan}@icmc.usp.br
http://www.nilc.icmc.usp.br

Abstract. The fulfillment of the UNL vision is primarily conditioned on the successful deployment of deconverters, each translating from the UNL into a target language. According to current practice, developing deconverters ultimately means configuring DeCo, the deconversion engine provided by the UNDL Foundation. However, DeCo has a number of limitations that hinder productivity and might even preclude quality deconversion. This paper discusses some of these shortcomings and introduces an alternative deconversion model – Manati, which is the result of work on UNL-mediated Portuguese-Brazilian Sign Language human-aided machine translation. With Manati we attempt to exemplify how multiparadigm – namely, constraint, object-oriented and higher-order – programming can be drawn upon not only to specify an open-architecture, optimum-searching deconversion engine but also and above all to rationalize its configuration into deconverters for target languages.

1 Introduction

The fulfillment of the UNL vision [10, 11, 18] is primarily conditioned on the successful deployment of deconverters, each translating from the UNL into a target language. UNL deconversion is actually an instance of Natural Language Generation (NLG), which refers to rendering linguistic form to input in a non-linguistic representation. As pointed out by e.g. Reiter & Dale [13], Cahill & Reape [3], and Paiva [12], NLG can be a very complex task involving processing both linguistic (e.g. lexicalization, aggregation and referring expression generation) and otherwise (e.g. content selection and layout planning). The good news is that UNL deconversion is in fact restricted to the linguistic aspect of NLG, which can be termed **linguistic realization** and comprises the usual macro-level tasks of microplanning and surface realization. Therefore, one should naturally expect UNL deconversion to benefit from recent advances in Natural Language Generation and software development practice, for which reason UNL developers may need to go beyond the model underlying the DeConverter – or simply DeCo, the generic deconversion engine provided by the UNDL foundation.

In this paper we analyze DeCo both as a formal object and a software product, with an emphasis on discussing DeCo's features that may hinder productivity. In this analysis we adopt configurability (i.e. ease of configuration into full-fledged decon-

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 175–194.

Arabic Generation
in the Framework of the Universal Networking Language

Daoud Maher Daoud

ALzaytoonah University, Amman, Jordan
daoud_m@yahoo.com , daoud@maherinfo.com
&
GETA, CLIPS, IMAG
daoud.daoud@imag.fr

Abstract. This paper describes the work done on the developing of Arabic Deconversion within the framework of the Universal Networking Language (UNL). In this paper, the architecture of the system is explained along with the strategy used for the development. We also discuss issues and problems related to the UNL representation that affect the quality of generation. Additionally, the lingware engineering is introduced as a technique to enhance the quality and increase the development efficiency.

1 Introduction

Arabic is one of the world's main languages. It is the official language for over 289 million people. It is also the sacred language of nearly 1.48 billion Muslims throughout the world.

The alphabet consists of twenty-eight consonants but three of these are used as long vowels. Arabic also contains short vowel signs being indicated by marks above or below the letters. Like other Semitic languages, Arabic is written from right to left. It is a language characterized by rich morphology: most of the words are built from consonantal roots in which inflections and derivations are generated by vowel changes, insertions, and deletions.

The Universal Networking Language is a specification for the exchange of information. It is a formal language for symbolizing the sense of natural language sentences.

Currently, the UNL includes 16 languages. These include the six official languages of the United Nations (Arabic, Chinese, English, French, Russian and Spanish), in addition to ten other widely spoken languages (German, Hindi, Italian, Indonesian, Japanese, Latvian, Mongol, Portuguese, Swahili and Thai). In its second phase (1999–2005) the project will seek to further extend UNL access.

This paper presents the work completed on the generation of Arabic from UNL during the author’s employment with Royal Scientific Society (RSS) in Jordan and his work on the UNL project. It described the work done on the generation of Arabic from UNL between 1996 till 1999. Since then, we think that the generation system maintained its main architecture.

Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 195–209.

Development of the User Interface Tools for Creation of National Language Modules

Tigran Grigoryan, Vahan Avetisyan

Institute for Informatics and Automation Problems
P. Sevak Str. 1,
375014 Yerevan, Armenia
{va@ipia.sci.am; tigrangr@ipia.sci.am }
http://ipia.sci.am; http://www.unl.am

Abstract. The paper describes the UNL Toolbox, software for development of national language modules of UNL, designed at the Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia. The software provides tools for creating dictionaries, enconversion and deconversion rules. There are also enconversion and deconversion modules which output the converted text and the list of occurred errors and their descriptions (if any). This software also can be used as an educational tool to learn creating UNL dictionaries and conversion rules.

1 System Overview

The UNL Toolbox is an integrated environment for UNL development. It contains tools for performing the most common tasks arising during UNL development such as dictionary creation and conversion rules creation. The Toolbox makes routine tasks like compilation of dictionary and conversion rules transparent to the end user. It allows setting options for an individual tool as well as for a whole system (for example the common output directory).

The main window of the Toolbox is divided into two parts (Fig. 1). On the left side of the window the toolbar is located. Pressing the buttons on the toolbar brings up appropriate tool in the right side of the window. Currently four tools are available – Dictionary Editor, Enconversion Rules Editor, Deconversion Rules Editor and Converter.

2 Dictionary Editor

Dictionary Editor provides a user friendly interface for creating UNL dictionaries and editing the existing ones. It uses XML to store the dictionary. When needed it is possible to export the dictionary in a standard plain text UNL dictionary format.

The dictionary in the dictionary editor has a tree-like structure. Each word is represented as a node of a tree with its stems (if any) represented as child nodes (Fig. 2).

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 210–214.

Universal Networking Language Based Analysis and Generation for Bengali Case Structure Constructs

Kuntal Dey[1] and Pushpak Bhattacharyya[2]

[1] Veritas Software, Pune, India.
u2ckuntal@yahoo.com
[2] Computer Science and Engineering Department
Indian Institute of Technology, Bombay, India.
pb@cse.iitb.ac.in

Abstract. Case structure analysis forms the foundation for any natural language processing task. In this paper we present the computational analysis of the complex case structure of Bengali- a member of the Indo Aryan family of languages- with a view toward interlingua based MT. Bengali is ranked 4^{th} in the list of languages ordered according to the size of the population that speaks the language. Extremely interesting language phenomena involving morphology, case structure, word order and word senses makes the processing of Bengali a worthwhile and challenging proposition. A recently proposed scheme called the *Universal Networking Language* has been used as the interlingua. The approach is adaptable to other members of the vast Indo Aryan language family. The parallel development of both the analyzer and the generator system leads to an insightful intra-system verification process in place. Our approach is *rule based* and makes use of authoritative treatises on Bengali grammar.

1 Introduction

Bengali is spoken by about 189 million people and is ranked 4^{th} in the world in terms of the number of people speaking the language (ref: *http://www.harpercollege.edu/˜mhealy/g101ilec/intro/clt/cltclt/top100.html*). Like most languages in the Indo Aryan family, descended from Sanskrit, Bengali has the SOV structure with some typical characteristics. A motivating factor for creating a system for processing Bengali is the possibility of laying the framework for processing many other Indian languages too.

Work on Indian language processing abounds. *Project Anubaad* [1] for machine translation from English to Bengali in the newspaper domain uses the *direct translation approach. Angalabharati* [2] system for English Hindi machine translation is based on pattern directed rules for English, which generates a *pseudo-target-language* applicable to a group of Indian Languages. In MATRA [3], a web based MT system for English to Hindi in the newspaper domain, the input text is transformed into case-frame like structures and the the target

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 215–229.

Interactive Enconversion by Means of the Etap-3 System

Igor M. Boguslavsky, Leonid L. Iomdin and Victor G. Sizov

Institute for Information Transmission Problems RAS, 19, Bolshoj Karetnyj, GSP-4
Moscow, Russia,
{bogus,leonid,sizov}@iitp.ru

Abstract. A module for enconversion of NL texts into Universal networking Language (UNL) graphs is considered. This module is designed for the system of multi-lingual communication in the Internet that is being developed by research centers of about 15 countries under the aegis of UN. The enconversion of NL texts into UNL is carried out by means of a multi-functional linguistic processor ETAP-3, developed in the Computational linguistics laboratory of the Institute for Information Transmission Problems of the Russian Academy of Sciences. One of the major problems in the automatic text analysis is high degree of ambiguity of linguistic units. The resolution of this ambiguity (morphological, syntactic, lexical, translational) is partly ensured by the linguistic knowledge base of ETAP-3, but complete algorithmic solution of this problem is unfeasible. We describe an interactive system that helps resolve difficult cases of linguistic ambiguity by means of a dialogue with the human.

1 Introductory Remarks

ETAP-3 is a multipurpose NLP environment that was conceived in the 1980s and has been worked out in the Institute for Information Transmission Problems, Russian Academy of Sciences ([1], [2], [7]). The theoretical foundation of ETAP-3 is the Meaning ⇔ Text linguistic model by Igor' Mel'čuk and the Integral Theory of Language by Jurij Apresjan. ETAP-3 is a non-commercial environment primarily oriented at linguistic research rather than creating a marketable software product. The main focus of the research carried out with ETAP-3 is computational modelling of natural languages. All NLP applications in ETAP-3 are largely based on a three-value logic and use an original formal language of linguistic descriptions, FORET.

2 Briefly on ETAP-3

The major NLP modules of ETAP-3 are as follows:

- Machine Translation System
- Natural Language Interface to SQL Type Databases
- System of Synonymous Paraphrasing of Sentences
- Syntactic Error Correction Tool

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 230–240.

Prepositional Phrase Attachment and Interlingua

Rajat Kumar Mohanty, Ashish Francis Almeida, and Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai - 400076, India
Emails: {rkm, ashishfa, pb}@cse.iitb.ac.in

Abstract. In this paper, we present our work on the classical problem of *prepositional phrase attachment.* This forms part of an interlingua based machine translation system, in which the semantics of the source language sentences is captured in the form of *Universal Networking Language (UNL)* expressions. We begin with a thorough linguistic analysis of six common prepositions in English, namely*, for, from, in, on, to* and *with.* The insights obtained are used to enrich a *lexicon* and a *rule base*, which guide the search for the correct attachment site for the prepositional phrase and the subsequent generation of accurate semantic relations. The system has been tested on British National Corpus, and the accuracy of the results establishes the effectiveness of our approach.

1 Introduction

No natural language processing system can do a meaningful job of analyzing the text, without resolving the prepositional phrase (PP) attachment. There are two fundamental questions related to this problem:

(1) *Given a sentence containing the frame*
[V-NP$_1$-P- NP$_2$]
does NP$_2$ attach to V or to NP$_1$?
(2) *What should be the semantic relation that*
links the PP with the rest of the concept graph of the sentence?

Our work is motivated by seeking answers to these questions. We focus our attention on six most common prepositions of English, *viz., for, from, in, on, to* and *with* (for the motivation, please see Table 5 in section 5).

In order to resolve these issues, we have taken linguistic insights from the following works [1–4]. Other related and motivating works specific to the PP-attachment problem are [5–9].

The roadmap of the paper is as follows: Section 2 provides a linguistic analysis of the six prepositions in question. The UNL system is introduced in Section 3. Section 4 discusses the design and implementation of the system. Evaluation results are given in Section 5. Section 6 concludes the paper and is followed by the references.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 241–253.

Hermeto: A NL-UNL Enconverting Environment

Ronaldo Martins, Ricardo Hasegawa and M. Graças V. Nunes

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação
Av. do Trabalhador São-Carlense, 400. CEP 13560-970. São Carlos – SP – Brasil
ronaldo@nilc.icmc.usp.br; gracan@icmc.usp.br; rh@nilcicmc.usp.br
http://www.nilc.icmc.usp.br

Abstract. This paper aims at presenting and describing HERMETO, a computational environment for fully-automatic, both syntactic and semantic, natural language analysis. HERMETO converts a list structure into a network structure, and can be used to enconvert from any natural language into the Universal Networking Language (UNL). As a language-independent platform, HERMETO should be parameterized for each language, in a way very close to the one required by the UNL Center's EnConverter. However, HERMETO brings together three special distinctive features: 1) it takes rather high-level syntactic and semantic grammars; 2) its dictionaries support attribute-value pair assignments; and 3) its user-friendly interface comprises debug, compiling and editing facilities. In this sense, HERMETO is said to provide a better environment for the automatic production of UNL expressions.

1 Introduction

In the UNL System [1], natural language (automatic) analysis has been carried out either by the EnConverter (EnCo) [2] or, more recently, by the Universal Parser (UP) [3], both provided by the UNL Center. In the first case, enconverting from natural language (NL) to Universal Networking Language (UNL) is supposed to be conducted in a fully-automatic way, whereas in the second case a full-fledged human tagging of the input text should be carried out before NL analysis is triggered. In both cases, results have not been adequate. EnCo's grammar formalism, as well as UP's tagging needs, are rather low-level, and requires a human expertise seldom available. In what follows, we present an alternative analysis system, HERMETO, developed at the Interinstitutional Center for Computational Linguistics (NILC), in Sao Carlos, Brazil, which has been used for automatic enconverting from English and Brazilian Portuguese into UNL. Due to its interface debugging and editing facilities, along with its high-level syntactic and semantic grammar and its dictionary structure, it is claimed that HERMETO may provide a more user-friendly environment for the production of UNL expressions than EnCo and UP.

The structure of this paper is as follows. The second section, on motivation, addresses the context in which the HERMETO initiative was conceived and the goals ascribed to the system. The third section presents HERMETO's architecture. HERMETO's functioning is briefly detailed in section four (on resources) and five

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 254–260.

A Platform for Experimenting with UNL

Wang-Ju Tsai

GETA, CLIPS-IMAG
BP53, F-38041 Grenoble cedex 09 France
Wang-Ju.Tsai@imag.fr

Abstract. We introduce an integrated environment, which provides the initiation, information, validation, experimentation, and research on UNL. This platform is based on a web site, which means any user can have access to it from anywhere. Also we propose an XML form of UNL document as the base of future implementation of UNL on the Internet.

1 Introduction

Since proposed 5 years ago, UNL project has attracted 16 international teams to join and is regarded as a very promising semantic Interlingua for knowledge representation on the Internet. The articles and applications of UNL have been found in many domains such as: machine translation, information retrieval, multilingual document generation, etc. Now we can find on the Internet not only the web sites of UNL language centres but also some discussions. The applications to facilitate the usage of UNL have been produced as well. Now we see the need to create a platform to integrate these applications also to introduce UNL to new ordinary users. We create this platform on a web site SWIIVRE (http://www-clips.imag.fr/geta/User/wang-ju.tsai/welcome.html), which has several goals: for the initiation, information, verification, research, and experimentation of UNL. And since this platform is based on a web site, any user from anywhere can have access to it.

2 Introduction of the Site SWIIVRE

In Appendix I we list all the resources accessible for UNL society members from internet. We can find out that most of the LC's connect vertically to UNL Centre but the horizontal connection among LC's is not enough, which means any user who wants to try the multilingualism of UNL will feel frustrated, since he will need to spend a lot of time try out every LC to know what service he can get.

The main purpose of this site is rather to integrate the current UNL applications and complete the services of Language Centres', when the function is available on a Language Centre, we simply provide the link to it, we also produce some applications to integrate or provide new functions, which all serve to facilitate the usage of UNL. Also we collect the useful information and publications on UNL, the web site is updated regularly. Lastly, by collecting the useful information and recording the related

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 261–267.

A Framework for the Development of Universal Networking Language E-Learning User Interfaces

Alejandro Martins,[1] Gabriela Tissiani,[2] and Ricardo Miranda Barcia [3]

[1] Distance Learning Laboratory, Federal University of Santa Catarina,
Campus da Trindade, Florianópolis, SC, Brazil,
martins@led.br

[2] CNPq Researcher at Centre Universitaire d'Informatique,
University of Geneva, Switzerland,
Gabriela.Tissiani@cui.unige.ch

[3] Distance Learning Laboratory, Federal University of Santa Catarina,
Campus da Trindade, Florianópolis, SC, Brazil
rbarcia@eps.ufsc.br

Abstract. The UNL infrastructure aims to overcome the language barrier on the Internet. At the same time, distance learning (DL) is becoming the best way to promote the knowledge diffusion across countries. However, the distance learning process still presents some obstacles to be overcome. The UNL can help to reduce particularly those problems and to provide a common educational environment across different languages. Here we discuss the development of the UNL version of an existing web platform for distance learning. The overall goal of this project is to create a framework to support the development of UNL user interfaces applied for e-learning platforms.

1 Introduction

This paper presents part of a research project that aims to build an e-learning platform using Universal Networking Language (UNL) technology.

It envolves the prototype development of the UNL version of an existing e-learning platform called VIAS-K (Virtual Institute of Advanced Studies - Knowledge Environment). This platform is provided by the Distance Teaching Laboratory, LED, from the Federal University of Santa Catarina, UFSC, Brazil [15]. It supports a huge group of interactive models composed of actors, contents, management, users support and collaborative tools. In order to full fill each user's specific needs, theses models will consider also the variety of users' mother language, based on the UNL system.

Although UNL have been developed with success, it is still a brand-new technology [17]. It is an artificial language that exchanges the knowledge from a natural language to make possible the access of its content through different languages. With the purpose of promoting the development of UNL and the effectiveness of DL, this research proposes a case study that brings UNL into the existing VIAS-K environment.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 268–275.

A WEB Platform Using UNL: CELTA’s Showcase

Lumar Bértoli Jr.,[1] Rodolfo Pinto da Luz,[2] and Rogério Cid Bastos [3]

[1] Instituto UNDL Brasil, Rodovia SC 401 - Km 1 –
ParqTec Alfa, Ed. CELTA - Bloco B - 3º pavimento - Módulo 2.10
João Paulo – 88030-000, Florianópolis– SC- Brasil
bertoli@undl.org.br

[2] Instituto UNDL Brasil, Rodovia SC 401 - Km 1 –
ParqTec Alfa,Ed. CELTA - Bloco B - 3º pavimento - Módulo 2.10
João Paulo – 88030-000, Florianópolis– SC- Brasil
luz@undl.org.br

[3] Rogério Cid Bastos, Departamento de Informática e Estatística, UFSC,
Campus Universitário - Trindade - 88040-900, Florianópolis - SC – Brasil
rogerio@inf.ufsc.br

Abstract. Economic globalization is changing the way companies communicate. The ease and speed of accessing information and taking decisions is better for everyone on the decision chain. To speed up access to information it is important to present information in one’s native language and create a language-independent communications channel. To assist business-to-business operations, the development team at the Instituto UNDL Brasil designed the pilot project “CELTA's Showcase” to demonstrate that it is possible to create a multilingual business-to-business platform using UNL.

1 Introduction

The interconnection between producer and consumer is becoming extremely important. The expansion of markets from local to global influence requires the use of new technological resources in order to support the majority of these relationships. In addition, the specialization of markets requires the development of automated tools to facilitate the pairing of small groups of producers and consumers.

Due to the irreversible globalization of markets and the specialization of production areas that create high technology products, there is a growing need for perfect matching between producers and consumers, to allow maximum performance in efforts to connect both sides.

To increase the chances of matching the best producer-consumer pair, the *Instituto UNDL Brasil is* proposing a project in this field. The main objective of this project is the development of a multilingual Web platform that allows integration between producing companies and their customers. This project is being developed by the *Institute UNDL Brasil*, and was made possible by the creation of the *UNL Research and Development group (R&D)* in the year 2003 [1]. The R&D group has a highly trained IT team whose main objectives are:

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 276–285.

Studies of Emotional Expressions in Oral Dialogues: towards an Extension of Universal Networking Language

Mutsuko Tomokiyo,[1] Gérard Chollet,[2] Solange Hollard [3]

[1] GETA-CLIPS-IMAG & ENST
BP 53 38041 Grenoble cedex 9 France
mutsuko.tomokiyo@imag.fr ; tomokiyo@tsi.enst.fr
[2] ENST
46 rue Barrault, 75634, Paris
chollet@tsi.enst.fr
[3] GEOD-CLIPS-IMAG
BP 53 38041 Grenoble cedex 9 France
Solange.Hollard@imag.fr

Abstract. Emotions entail distinctive ways of perceiving and assessing situations, processing information, and prioritizing and modulating actions [24]. The paper aims to study theoretical and pragmatic aspects of emotions and to propose a semantic representation of emotions for oral dialogues, based on an analysis of real-life conversations, telephone messages and recorded TV programmes, focusing on a relationship between prosody and lexeme for the purposes of a speech to speech machine translation. The semantic representation is made, by using the **U**niversal **N**etworking **L**anguage (UNL) formalism, in a way where lexeme, phatics, gestures, prosody and voice tone are taken into account at the same time.

1 Introduction

This work has been carried out in a continuation of “VoiceUNL” [21], which is one of subprojects of the “LingTour” [1] project. “VoiceUNL” is an extension of **U**niversal **N**etworking **L**anguage (UNL), which is a text-oriented formalism of semantic graphs,, to oral dialogues.

As for **s**peech to **s**peech **m**achine **t**ranslations (SSMT) or man-machine interactive systems, the detection and generation of emotions are an important issue from the viewpoint of the naturalness of dialogues [7], because *emotion entails distinctive ways of perceiving and assessing situations, processing information, and prioritizing and modulating actions* [24]. It's the key reason for proposing a semantic representation of emotions.

In this paper, section 2 is devoted to previous emotion studies mainly focussed on prosody: a survey of existing approaches to emotion detection and generation, theo-

[1] The Lingtour project was launched in 2002 by the partnership which consists of TsingHua University (China), Paris 8 University (France), INT (France), ENST-Paris and Bretagne (France), and CLIPS (France). One of the objectives of the projects resides in R & D to enable multilingual-multimedia MT on user-friendly tools [1].

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 286–299.

An XML-UNL Model for Knowledge-Based Annotation

Jesús Cardeñosa, Carolina Gallardo and Luis Iraola

Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Madrid, Spain
{carde,carolina,luis}@opera.dia.fi.upm.es

Abstract. Efficient document search and description has radically changed with the widespread availability of electronic documents through Internet. Nowadays, efficient information search systems require to go beyond HTML-annotated documents. Complex information extraction tasks require to enrich text with semantic annotations that allow deeper and more detailed content analysis. For that purpose, new labels or annotations need to be defined. In this paper we propose to use UNL, an interlingua defined by the United Nations University, as a language neutral standard content representation in Internet. The use of UNL would open documents to a new dimension of semantic analysis, thus overcoming the limitations of current text-based analysis techniques.

1 Introduction

XML [1] is an standardized annotation language currently employed for a variety of purposes. For any given domain, the set of tags defined in its DTD attempts to capture the logical content structure of typical documents of the domain. So annotated, documents can be exploited by sophisticated document management systems that provide precise answers to users' queries. One the most promising uses of XML is the possibility of replacing textual document bases by their XML counterparts for document management purposes as well as for content management.

The capability of the XML standard to define the different information items present in a given document facilitates subsequent information extraction operations. This capability makes XML an ideal choice for annotating text corpora.

Annotated corpora have been one of the most useful resources in the last years for the study of linguistic phenomena. This orientation towards linguistic analysis has frequently associated corpus annotation with tasks such as part of speech tagging, chunking and parsing.. The Brown Corpus [2] or the British National Corpus [3] are examples of such annotated corpora. This sort of annotation is useful for many purposes but may be insufficient for information management tasks and for the location of very specific information items.

Corpus annotation poses significant difficulties when the goal is the representation and classification of information expressed in text form. While one could say that lexical and syntactic annotation of textual corpora is a more or less solved problem, semantic tagging is still a challenging goal currently aimed by several research lines.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 300–308.

A "Pivot" XML-Based Architecture for Multilingual, Multiversion Documents: Parallel Monolingual Documents Aligned Through a Central Correspondence Descriptor and Possible Use of UNL

Najeh Hajlaoui, Christian Boitet

équipe GETA, laboratoire CLIPS,
385 rue de la bibliothèque - BP 53, 38041 Grenoble Cedex 9 - France
Christian.Boitet@imag.fr

Abstract. We propose a structure for multilingual, multiversion documents, built on the model of the web-oriented, cooperative lexical multilingual data base PAPILLON: a document is represented by a collection of monolingual XML "volumes" interlinked by a central volume of "interlingual links". Here, the links relate subdocuments (XML trees) corresponding to each other in monolingual "volumes". We are developing a Java application to enable direct editing of a multilingual document through the web, at the level of monolingual volumes as well as through bilingual or trilingual interfaces inspired by those of commercial "translation workbenches". Another goal is easy integration with machine translation and multilingual generation tools. For this, we add a special UNL volume. In a first stage, we split the UNL-xml document in several monolingual documents, again represented by XML files. Each document contains the text in a particular language, plus the corresponding UNL graphs, and can be modified independently. The interface is easy to build, but realigning the documents after a series of such modifications is a very difficult task.

1 Introduction

Due to Internet, the number of available documents grows dramatically. There is a strategic need for companies to control information written in more than 30 languages (HP, IBM, MS, Caterpillar). This requires the installation of powerful and effective management tools of multilingual "synchronized" documents.

There are techniques of large-grained linking (on the level of HTML pages). However, there are no techniques for structuring multilingual documents so as to allow fine-grained synchronization (at paragraph or sentence level) and even less permitting editability through the Web.

The interest to synchronize at least on the level of the sentences is double:

- for the translation and human revision with the assistance of techniques of HTHM (Human Translation Helped by Machine) and in particular of translation memory.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 309–325.

UCL—Universal Communication Language

Carlos A. Estombelo-Montesco and Dilvan A. Moreira

Universidade de São Paulo,
Instituto de Ciências Matemáticas e de Computação
Av. do Trabalhador São-Carlense, 400 - Centro - Cx. Postal 668
São Carlos - SP - Brazil CEP 13560-970
c_estombelo@yahoo.com, dilvan@computer.org

Abstract. For successful cooperation to occur between agents they have to be able to communicate among themselves. To enable this communication an Agent Communication Language (ACL) is required. Messages coded in an ACL should adequately express their meaning from a semantic point of view. The Universal Communication Language (UCL) can fulfill the role of an ACL and, at the same time, be convertible to and from a natural language. UCL design is concerned with the description of message structures, their underlining semantic context and the support for protocols for agent interaction. The key point about UCL is that the language can be used not only for communication among software agents but among humans too. This is possible because UCL is derived from the Universal Network Language (UNL), a language created to allow communication among people using different languages. UCL was defined using the Extended Markup Language (XML) to make it easier to integrate into the Internet. In addition, an enconverter-deconverter software prototype was written to serve as a tool for testing and experimenting with the language specifications.

1 Introduction

The technology of software agents can be an interesting tool for the creation of new models for complex software systems. In the project of software agents, many of the traditional techniques of artificial intelligence can be mixed with techniques from the field of distributed computer systems, theories about negotiation and theories about working teams [2]. Software agents are basically designed to cooperate (either with others or with humans) in a seemingly intelligent way. But for cooperation to occur a communication language is necessary.

What does it mean to be able to communicate with someone? Simplifying it, useful communication requires shared knowledge. While this includes knowledge of language, words and syntactic structures, meaningful communication is even more focused on knowledge about a problem to be solved. To interact with a florist you need some knowledge about flowers.

The widespread use of the Word Wide Web (WWW) and growing Internet facilities have sparked enormous interest in improving the way people communicate using computers. To date, communication among software agents and

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 326–336.

Knowledge Engineering Suite: A Tool to Create Ontologies for Automatic Knowledge Representation in Intelligent Systems

Tânia C. D. Bueno,[1] Hugo C. Hoeschl,[1] Andre Bortolon,[1] Eduardo S. Mattos,[1]
Cristina Santos,[1] Ricardo M. Barcia [2]

[1] Instituto de Governo Eletrônico, Inteligência Jurídica e Sistemas – IJURIS
Rua Lauro Linhares, 728, sala 105, Florianópolis BRASIL – CEP 88036-002
http://www.ijuris.org
tania@ijuris.org;{hugo, bortolon, mattos, cristina}@wbsa.com.br

[2] Virtual Institute of Advance Studies – VIAS
Florianópolis, BRASIL
rbarcia@uol.com.br

Abstract. The present work is focused on the systematization of a process of knowledge acquisition for its use in intelligent management systems. The result was the construction of a computational structure for use inside the institutions (Intranet) as well as outside them (Internet). This structure was called Knowledge Engineering Suite, an ontological engineering tool to support the construction of ontologies in a collaborative environment and was based on observations made at Semantic Web, UNL (Universal Networking Language) and WordNet. We use both a knowledge representation technique called DCKR to organize knowledge, and psychoanalytic studies, focused mainly on Lacan and his language theory to develop a methodology called Engineering of Mind to improve the synchronicity between knowledge engineers and specialists in a particular knowledge domain.

1 Introduction

The importance of the Knowledge Based Systems is in the fact that they provide the computer with some peculiar characteristics of human intelligence, such as the capacity to understand natural language and simulate reasoning in uncertainty conditions. Defining the relevant information to be inserted into a Knowledge Based Systems is the great problem in the development of intelligent systems, mainly because the process is basically experimental and depends greatly on the ability of the knowledge engineer. In particular, a great difficulty is related to the definition of the terminology used to nominate the concepts and the relations [1]. Besides the great number of methods to do the knowledge acquisition, we can't find one that deals with the understanding and learning of the people involved, both specialists and knowledge engineers.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 337–346.

Using Semantic Information to Improve Case Retrieval in Case-Based Reasoning Systems

J. Akshay Iyer and Pushpak Bhattacharyya

Indian Institute of Technology Bombay
{akshay,pb}@cse.iitb.ac.in

Abstract. Conventional Case-Based Reasoning (CBR) systems rely on word knowledge to index and search cases from its memory. On being presented with a problem, the Case-Based Reasoning system tries to retrieve a relevant case based on the words that appear in the problem sentence without considering their respective senses. Drawbacks of such systems become more evident in cases where the input is in the form of a sentence in a natural language. Ignoring semantic information in this case may not result in retrieval of desired case or may result in retrieval of an undesired case. In this paper we present a method that tries to improve the precision of retrieval by also taking into account semantic information available to us about the words in the problem sentence. Towards this goal, Universal Networking Language (UNL) is made use of, which provides a semantic representation of natural language text to capture sentence structure. Lexical resource like WordNet is used for finding semantic similarity between two concepts. Using an existing commercial Case-Based Reasoning system as basis for comparison, we demonstrate that considering such semantic information helps in improving case retrieval.

1 Introduction

Case-Based Reasoning (CBR) Systems are one of the most widely used systems in the field of problem solving and planning. A number of such systems are developed and reported [5, 8, 11]. Typically, a number of cases are stored in memory and upon being presented with a problem, a set of relevant cases is retrieved and presented as a solution to the problem [7]. One of the fundamental issues in such systems concerns this retrieval process. Information from the input problem is extracted out and this information is used to index (or search) in the memory to locate the desired case. In systems where a problem is input in Natural Language form, the issue becomes more profound. Traditionally, a number of statistical methods are used for extracting information from the input problem and using it in turn for identifying cases that are relevant to the problem. However, since such methods do not employ any natural language understanding, they fail in situations when mere knowledge about words is not sufficient.

In this paper we propose a method by which we could use information, both semantic and syntactic, from natural language text to compare and retrieve relevant cases. The rest of the paper is organized as follows. Section 2 discusses the shortcom-

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 347–358.

Facilitating Communication Between Languages and Cultures: a Computerized Interface and Knowledge Base

Claire-Lise Mottaz Jiang,[1] Gabriela Tissiani,[2]
Gilles Falquet,[1] Rodolfo Pinto da Luz [3]

[1] CUI, University of Geneva, Switzerland,
(Claire-Lise.Mottaz, Gilles.Falquet)@cui.unige.ch
[2] CNPq Researcher at CUI, University of Geneva, Switzerland
Gabriela.Tissiani@cui.unige.ch
[3] Instituto UNDL Brasil, Florianópolis, Brazil
Luz@undl.org.br

The Universal Networking Language (UNL) deals with communication, information, knowledge, language, epistemology, computer sciences, and related disciplines. This interdisciplinary endeavor calls for theoretical and applied research, which can result in a number of practical applications in most domains of human activities. Specially, it can help solving some of the most critical problems emerging from current globalization trends of markets and geopolitical interdependence among nations. This paper presents a project that aims to contribute with UNL KB (UNL Knowledge Base) theoretical and practical. The goal is to make possible people from various linguistic and cultural backgrounds to participate at UNL KB construction in a distributed environment.

1 Introduction

This paper presents a project that will be developed by the following partners: Information System Interfaces (ISI) Research Group at University of Geneva, the UNDL Foundation, and the United Nations Institute for Training and Research (UNITAR). This project is part of the Geneva International Academic Network Programme (GIAN). It involves creation of ontologies, for the Universal Networking Language (UNL) Knowledge Base (KB).

The project argues that the construction of these KB ontologies will contribute to the United Nations initiative of creating the multilingual infrastructure on UNL. Its infrastructure is meant to facilitate communication among natural languages on the Internet and includes development of a broad knowledge base from diverse linguistic sources and cultural backgrounds [10].

The UNL multilingual infrastructure is an interdisciplinary undertaking that involves both linguistic and engineering aspects. Its main components are (1) a formal, language-independent, non-ambiguous artificial language (UNL) and (2) a system that manages the interfaces between natural languages and the UNL over computer networks. The UNL itself comprises a vocabulary - a list of concepts, called "univer-

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 359–369.

Using WordNet for linking UWs to the UNL UW System

Luis Iraola

Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Madrid, Spain
luis@opera.dia.fi.upm.es

Abstract. This paper presents the work done with the Spanish-UNL dictionary compiled at the Spanish Language Centre in order to enrich the universal words it contained with the supplementary semantic information required to produce a master entries dictionary. Focusing on a subset of the Spanish-UNL dictionary, namely on the substantives it contains, the work has consisted in automatically enrich the universal word associated with each substantive with the semantic information required to link the universal word to the Universal Word System. For this process, WordNet has been employed as an external source of semantic information and used in addition to semantic features already present in the dictionary. The results achieved are not final and further work is required for a fully automatic, high quality semantic enrichment of the current entries. However, the work done shows the fruitfulness of the approach and its outcome has contributed to the creation of a master entries dictionary.

1 Introduction

A UNL dictionary in which language entries are associated with universal words (UW for short) can be viewed as a repository of UWs and as such does not organise its contents in any way. It links a set of UWs with lexical items of a specific language, each entry having no relation with any other. The necessity of establishing certain relations between UWs arises when considering several desirable features of the UNL system:

- Setting the combinatory possibilities of each UW with respect to any other UW regarding the conceptual relations that may link them and the attributes they may accept.
- Enabling a “fall-back” generation mechanism for those UWs that are not linked with head words in a given language at a given time. Those UWs would be replaced with semantically close but linked UWs so allowing generation to continue.

In order to support these features, a network with the set of UWs as nodes and semantic relations as arcs has been proposed. Such network is called the UNL UW System [1, 2]. Therefore, and in order to build the UW System, UNL Language Centres have to modify their respective UNL language dictionaries for including such new information. Once modified, the new master entries dictionaries will be the repository from which current language dictionaries will be produced as well as the UNL Knowledge Base will be created.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 370–379.

Automatic Generation of Multilingual Lexicon by Using Wordnet

Nitin Verma and Pushpak Bhattacharyya

Department of Computer Science and Engineering, I.I.T. Bombay,
{nitinv,pb}@iitb.ac.in

Abstract. A lexicon is the heart of any language processing system. Accurate words with grammatical and semantic attributes are essential or highly desirable for any application- be it machine translation, information extraction, various forms of tagging or text mining. However, good quality lexicons are difficult to construct requiring enormous amount of time and manpower. In this paper, we present a method for automatically generating multilingual Universal Word (UW) dictionaries (for English, Hindi and Marathi) from an input document-making use of English, Hindi and Marathi WordNets. The dictionary entries are in the form of Universal Words (UWs) which are language words (primarily English) concatenated with disambiguation information. The entries are associated with syntactic and semantic properties- most of which too are generated automatically. In addition to the WordNet, the system uses a word sense disambiguator, an inferencer and the knowledge base (KB) of the Universal Networking Language which is a recently proposed interlingua. The lexicon so constructed is sufficiently accurate and reduces the manual labor substantially.

1 Introduction

Construction of good quality lexicons enriched with syntactic and semantic properties for the words is time consuming and manpower intensive. Also word sense disambiguation presents a challenge to any language processing application, which can be posed as the following question: *given a document **D** and a word **W** therein, which sense **S** of **W** should be picked up from the lexicon?*. It is, however, a redeeming observation that a particular **W** in a given **D** is mostly used in a single sense throughout the document. This motivates the following problem: *can the task of disambiguation be relegated to the background before the actual application starts? In particular, can one construct a **Document Specific Dictionary*** wherein single senses of the words are stored?

Such a problem is relevant, for example, in a machine translation context [1]. For the input document in the source language, if the *document specific dictionary* is available a-priory, the generation of the target language document reduces to essentially syntax planning and morphology processing for the pair of languages involved. The WSD problem has been solved before the MT process starts, by putting in place a lexicon with the document specific senses of the words.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 380–391.

Gradable Quality Translations through Mutualization of Human Translation and Revision, and UNL-Based MT and Coedition

Christian Boitet

GETA, laboratoire CLIPS,
385 rue de la bibliothèque - BP 53, 38041 Grenoble Cedex 9, France
Christian.Boitet@imag.fr

Abstract. Translation of specialized information for end users into many languages is necessary, whether it concerns agriculture, health, etc. The quality of translations must be gradable, from poor for non-essential parts to very good for crucial parts, and translated segments should be accompanied with a measured and certified "quality level". We sketch an organization where this can be obtained through a combination of "mutualized" human work and automatic NLP techniques, using the UNL language of "anglosemantic" graphs as a "pivot". Building the necessary multilingual lexical data base can be done in a mutualized way, and all these functions should be integrated in a "Montaigne" environment allowing users to access information through a browser and to switch to translating or postediting and back.

1 Introduction

Translation of specialized information into many languages is necessary, notably in agriculture, but also for health and other domains, because it is often crucial for final users, who don't master the source language. Quality should be very high, at least for the crucial parts. In many cases, also, it is urgent to use the information, and only automated translation could offer a solution. At the same time, resources are scarce, especially to produce high quality translations. Does that mean that nothing can be done? No, of course.

The first idea which comes to mind is to "mutualize" the translation effort. That becomes possible thanks to the wide availability of Internet. There is always a minority of targeted readers who understand the source language, and could produce good translations. Also, they would translate only a fraction of their time, so that, even with machine helps which may be developed by and by, it is reasonable to assume that not every part of every document could be translated in this way. Why not, then, use "rough" machine translation (MT), or even "active reading helps" (annotations of the source text by possible translations of words, terms and even phrases), and have human readers decide on which crucial parts are difficult to understand when presented in this way, and improve them?

We claim that, in this and similar domains, the quality of translations theoretically can and practically must be gradable, from poor to very good. Translations of each

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 395–412.

Towards a Systematic Process in the use of UNL to Support Multilingual Services

Jesús Cardeñosa, Carolina Gallardo, Edmundo Tovar

Departamento de Inteligencia Artificial- Facultad de Informática
Universidad Politécnica de Madrid
Campo de Montegancedo, s/n
28660 Madrid- Spain
{carde, carolina, etovar}@opera.dia.fi.upm.es

Abstract. The UNL Programme of the United Nations University (UNU) was launched in 1996 aiming at the elimination of linguistic barriers in Internet. Now, eight years later, UNL is not ready to support real applications due to several circumstances. This eight-year period can be divided in two: a first four-year period devoted to the formal definition of UNL as a formal language (under the sponsorship of the Institute of Advanced Studies (IAS) of the UNU) and the remaining four years devoted to the technical experimentation of UNL. A new period is starting right now, which could be the period of maturity at all levels, especially at technical and business levels. In this paper, the authors summarize the more significant experiences until now, their conclusions and the set of procedures to produce marketable multilingual services. This kind of work will be the work of the UNL consortium during the next two years before launching UNL to the market.

1 Introduction

The natural evolution of UNL as a project and as a Programme is the support of useful applications for a multilingual society. Apart from other uses of UNL, like cross-lingual information retrieval or support for ontologies, the more understable use and possibly the easiest application, is the support of multilingual services, that is, to represent contents written in any language and to generate any other language [1].

UNL is not conceived to become a (fully automatic) machine translation system (MT hereafter). Up to date, MT systems based on the transfer architecture have achieved reasonable results, always involving pairs of languages. These systems are somehow handicapped by their *language coverage*. In other words, a transfer based system involving N languages requires the development of N × (N–1) systems, which ends up with the consequent combinatorial explosion of the number of systems to be developed as the number of languages grows.

On the other hand, interlingua-based MT systems show, in principle, a highly attractive advantage over transfer systems: interlingua-based systems do not grow exponentially as the number of language increases since for a system to support N languages, only 2 × N systems have to be developed. The ATLAS system [2] and the PIVOT system [3] in open domains, and Mikrokosmos [4] and Kant [5] in restricted

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 413–429.

Knowledge Representation Issues and Implementation of Lexical Data Bases

F. Sáenz and A. Vaquero

Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid,
E-28040 Madrid, Spain
Tel: +34913947622
Fax: +34913947529
{fernan,vaquero}@sip.ucm.es

Abstract. We propose to apply classical development methodologies to the design and implementation of Lexical Databases(LDB), which embody conceptual and linguistic knowledge. We represent the conceptual knowledge as an ontology, and the linguistic knowledge, which depends on each language, in lexicons. Our approach is based on a single language-independent ontology. Besides, we study some conceptual and linguistic requirements; in particular, meaning classifications in the ontology, focusing on taxonomies. We have followed a classical software development methodology for implementing lexical information systems in order to reach robust, maintainable, and integrateable relational databases (RDB) for storing the conceptual and linguistic knowledge.

1 Introduction

Due to the immaturity of the knowledge representation topic, lack of standardization is broadly felt as a very undesirable state into the community around language resources [LREC 02]. For instance, standard terminology for a common reference ontology is yet a goal to be reached. There is no doubt about what lexicon means, but ontology is differently understood in the computational linguistic literature. For instance, WordNet is mentioned as an ontology [USC 96], CYC is provided with a formal ontology [PRI 01], etc. Here, ontology, in a LDB, is the set of concepts in the domain of the base and the relationships that hold among them, without including linguistic knowledge, and common to all of the languages supported in the base.

Weak attention has been paid on topics about development methodologies for building the software systems which manage LDB, and dictionaries in particular. We claim that the software engineering methodology subject is necessary in order to develop, reuse and integrate the diverse available linguistic information resources. Really, a more or less automated incorporation of different lexical databases into a common information system, perhaps distributed, requires compatible software architectures and sound data management from the different databases to be integrated. The database subject have already done a long way reaching a strong standardization, and supplying models and methods suitable to develop robust information systems. We apply RDB design methodologies to develop LDB consisting of ontologies and

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 430–442.

Author Index

Índice de Autores