Multilingual Cross Language Information Retrieval
A new approach

Jesús Cardeñosa
Validation and Business Applications Research Group
Facultad de Informática – Universidad Politécnica de Madrid
Madrid – SPAIN
carde@fi.upm.es

Carolina Gallardo
Validation and Business Applications Research Group
E.U. de Informática – Universidad Politécnica de Madrid
Madrid – SPAIN
cgallardo@eui.upm.es

Adriana Toni
Validation and Business Applications Research Group
Facultad de Informática – Universidad Politécnica de Madrid
Madrid – SPAIN
atoni@fi.upm.es

ABSTRACT

CLIR is the acronym of a great variety of techniques, systems and technologies that associate information retrieval (normally from texts) in a multilingual environments. Many of these systems are based on a double architecture composed by systems in charge of extracting information with a great dependency on the language together with classical machine translation systems. In the early 90's, machine translation systems fell from grace due to the failure of big machine translations projects in Europe, Japan and USA. Due to this reason some approaches, particularly those of linguistic knowledge representation were undeservedly forgotten, and above all the so called "interlinguas". Recently, the re-emergence of these models under the generic name of "ontologies" are supporting most of knowledge representation initiatives, even in an language independent way However consistency problems are not well solved yet. UNL, initially conceived as a contents representation and multilingual generation system, can also be applied to the CLIR. This paper aims to discuss how the UNL could be considered as adequate for consistent knowledge representation in this type of systems.

Keywords

Information Retrieval, multilinguality.

1. INTRODUCTION

Cross-Language Information Retrieval (CLIR) deals with the problem of issuing a query in one language and retrieving relevant information in other languages. It aims to help the user in finding relevant information without being limited by linguistic barriers.

In order to overcome the language barrier, three major approaches exist:

- to translate the query into the documents' languages
- to translate the documents into the query's language
- to translate both into an intermediate representation through the use of domain-specific interlinguas.

1.1. Query Translation

Online translation can be applied to the query entered by the user. Online query translation will help the user to formulate his/her query in a language other than his/her own. If the user either has at least some reading skills in the target language, it may be possible for him/her to reformulate, elaborate or narrow down the translation proposed.

Because of its simplicity, query translation via machine-readable bilingual or multilingual dictionaries is a very most common approach [1]. Compared to translating an entire document collection, translating a query by dictionary look-up is far more efficient. However, it is unreliable since short queries do not provide enough context for disambiguation in choosing proper translations of query words, and also because it does not exploit domain-specific semantic constraints and corpus statistics in solving translation ambiguities.

A wide array of resources is used in CLIR [2], ranging from multilingual glossaries or dictionaries to multilingual collections of texts and sophisticated taggers and parsers (e.g., Mulinex [3] and MIETTA [4] projects).

1.2. Document translation

Full document translation can be applied offline to produce translations of an entire document. The translations provide the basis for constructing an index for information retrieval and also offer the user the possibility to access the content in his/her own language. Machine or (large scale) human translation, however, is not always available as a realistic option for every language pair. Typically machine translation systems only translate between language pairs which involve one of the major languages, such as English, German or Spanish, and often English plays a pivotal role.

1.3. Domain-specific ontologies for CLIR

Some representative CLIR projects (MuchMore [5], LIQUID [6]) employ a domain-specific ontology that contains the knowledge of the application domain and serves as an interlingual backbone for a multilingual thesaurus. Relevant terms contained in a query are translated into several languages using the term-to-concept links established in the multilingual thesaurus. Domain knowledge represented in the conceptual layer is exploited for expanding the initial query (see figure 1).

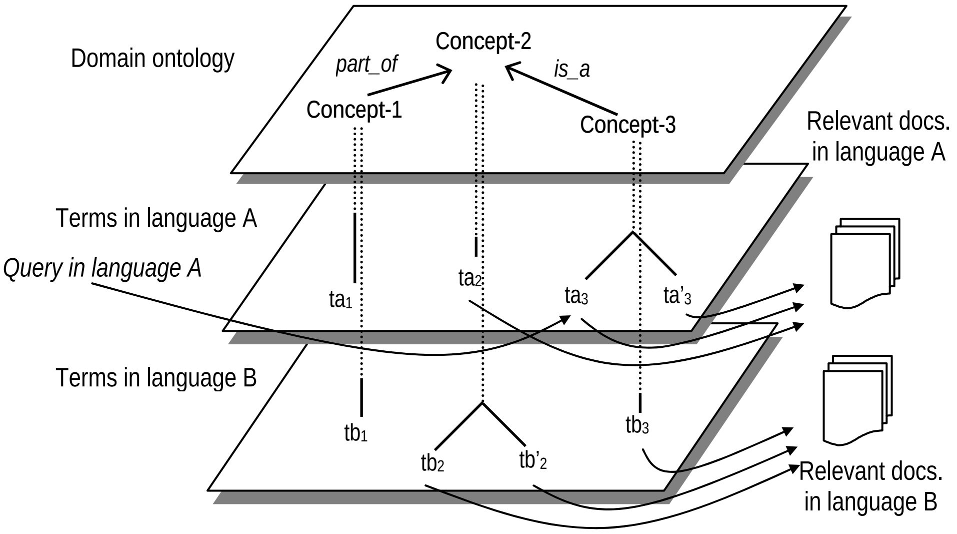

Fig 1: Linking documents and queries through a multilingually mapped ontology

2. ONTOLOGIES AND SUPPORT LANGUAGES

Like in many other cases, the definition of an ontology is not completely fixed and agreed on. There are several definitions of ontologies, but for our purpose we will cling to Gruber's one: "an ontology is an explicit specification of a conceptualization"[7].

There are two main issues in this definition:
a) Explicit specification
b) Conceptualization

The "explicit specification" of ontology leads us to the formalization of ontologies and used languages. In this section, we will deal with ontologies support languages as the main way for attaining such explicitness and machine readability.

A conceptualisation is related to the creation of a model of a given domain pointing out the relevant concepts, their relations and functions that made up a complete domain.

In order to support an ontology and inference mechanisms, the question of the language support is crucial. There are two main factors that determine the evolution of ontology languages. These are the *knowledge representation formalism* and *web orientation*.

Regarding the knowledge representation formalism, there appear to be two clear periods that we will refer to as *First Generation Languages* and *Second Generation Languages*. First generation ontology languages are basically frame-based and correspond to the first attempts to build ontologies and establish the ontology engineering discipline (beginning of 90ies). As the most representative frame-based languages are Loom [8], Ontolingua [9] or KIF [10].

In its beginnings, ontology engineering was highly oriented towards knowledge reuse and share [11]. All of these languages can be considered as languages for knowledge representation, being KIF (*Knowledge Interchange Format*) the most oriented towards knowledge reuse, since it conforms a sort of "interlingua" of knowledge representation languages.

The common feature of these languages is its frame-based nature. Thus, they are endowed with the usual expressiveness of frames. Basically, they allow for:

- Representing classes and subclasses
- Distinguishing between classes and instances
- Establishing relations between classes.
- Establishing default values.

In a way we could say that these languages are oriented toward a hierarchical conceptualisation of a domain. Needless to say, the Semantic Web wasn't the main goal in this period. So there is no web integration of these ontologies.

The second generation of ontology languages shows a more logical flavour (although some retain the frame flavour). We are referring to RDF [12], SHOE [13], DAML-OIL [14] or even XML [15]. Let's mention some of the properties of these languages:

- They are based on first order logic (with some possible extensions).
- Use of logic (formal semantics for deduction processes)
- The distinction between class and instance is supported.
- The establishment of taxonomies (class – subclass) is normally supported.
- Representation and inclusion of axioms are supported in some of them.
- Normally no default values are allowed.
- Relations (of different arity) are more or less covered.
- Some of them are oriented towards the Semantic Web (developed by the W3C consortium or either compatible with XML).

These ontology languages show the second parameter: web orientation, they extends the traditional definition of an ontology and try to conceptualise the whole web, that is, the target is no more reuse of knowledge but to achieve the so-called Semantic Web. Thus many of them are based on web languages and technologies (such as XML and RDF developed by the W3C consortium).

It is interesting to see the influence of an standard entity such as W3C as an standardizing body. It is quite obvious the convergence of all these languages towards a unique standard one: OWL [16].

All these languages seems to have derived in OWL, which is an extension of XML, RDF, DAML and OML. According to the authors, it provides “greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics”. It was in February, 2004 when it was proposed by W3C to become the standard language for ontology representations in the web.

3. KNOWLEDGE REPRESENTATION VS. CROSS-LINGUALITY

Ontologies and knowledge representation are two close concepts. At the end, conceptualisation and formalization of a model or domain are two quite well known issues of Knowledge representation. Historically, semantic nets was the first formalism suitable to represent knowledge, as it extended the expressiveness of pure logical models. The semantic nets were proposed in 1968 by Quillian and he was also who study the knowledge extraction from texts some years later [17]. Wood in [18] stated two issues that prevent semantic nets from being a good candidate for knowledge representation:

a) Ambiguities in its representation (no specific account of the distinction between class and instance)
b) Lack of a common understanding of the semantic labels, that eventually Wood defines as the “asemanticity” of semantic nets.

For these two reasons, ontology languages turn to frame and logic based formalisms, disregarding the adequacy of semantic nets for the specification of non-hierarchical relations (that is, functions and roles between concepts). Curiously, current ontologies do not fully exploit the most expressive characteristics of semantic nets, resulting in a massive use of relation IS-A. Bearing in mind the features of ontology languages, we could state that there is coverage for vertical relations (class, subclass, instance, plus other) but not for horizontal relations (roles and links between concepts). Horizontal relations enrich the domain representation, as shown in [19] and [20] as attempts to build ontologies from natural language texts. Even if we accept Wood’s objection to semantic nets, there is still a wide amount of information that semantic nets offers and ontologies do not exploit, being this the capacity of semantic nets to express horizontal relations, that could be easily integrated into ontology support languages in principle.

Thus relations would not be only limited to a is-a or a-kind-of types, but richer relations will have to be included. A hint of what sort of horizontal relation should be included in domain models is given by natural languages (languages are the main vehicle of expressing knowledge), this is the approach followed in the GUM, following the theoretical positions that Functional Grammar established [21], or as we will see later in the Universal Networking Language (UNL).

By knowledge bases in our context we understand the set of concepts belonging to a specific domain and the relations between these concepts that also belong to this domain. But when we turn to ontologies, the richness of a domain becomes relegated to a mere enumeration of concepts and a taxonomic organization of them. That is, there is danger of identifying ontologies as mere thesauri.

4. SOME ADVANCES: NEW APPROACHES

UNL is basically an artificial language for knowledge representation designed for representing contents written in any language and for generating such contents in any natural language. The next section will depict UNL in more detail.

4.1. UNL as interlingua

Formally speaking, UNL follows the schema of semantic nets (that is, UNL expresses binary relations between concepts, labelled by a number of semantic tags). The specifications of the language [22] formally define the set of relations, concepts and the so-called attributes. We will explain briefly the main ones, that is, the concepts and relations.

Universal words. They conform the vocabulary of the language. To be able to express any concept occurring in a natural language, the UNL proposes the use of English words modified by a series of semantic restrictions that eliminate the innate ambiguity of the vocabulary in natural languages. If there isn’t any English word suitable to express the concept, the UNL allows the use of words from other languages. In this way, the language gets an expressive richness from the natural languages but without their ambiguity.

Relations. These are a group of 41 relations that define the semantic relations among concepts. They include argumentative (agent, object, goal), circumstantial (purpose, time, place), logic (conjunction, and disjunction) relations, etc.

4.2. UNL as language for knowledge representation

UNL is mainly used as a support language for multilingual generation of contents coming from different languages. However, its design allows for non language centred applications, that is, UNL could serve as a support for knowledge representation in generic domains. When there is a need to construct domain-independent ontologies, researches turn back to natural language (such as Wordnet, GUM or even CyC[1]) to explore the “semantic atoms” that knowledge expressed in natural languages is composed of. UNL follows this philosophy, since it provides an interlingual analysis of natural language semantics.

But to really serve as a language for knowledge representation, it must support deduction mechanisms and must specify how a knowledge base could be build up in the UNL language. We will explore this idea by looking closer at the UWs part of the UNL system and how to link them in knowledge base.

4.3 The UNL dictionary and its companion KB

The UW dictionary is a repository of UWs and as such does not organise its contents in any way. It is just a (big) set of UWs, each element 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 HWs in a given language at a given time. Those UWs would be replaced with semantically close, linked UWs so allowing generation to continue.

If word sense disambiguation were the *only* reason for introducing semantic restrictions into UNL, any of the previous approaches could be adopted. The semantic restrictions attached to the UWs for disambiguation purposes *also* express knowledge stored in the KB and conversely; the semantic knowledge serves for disambiguation. Such network

[1] http://www.cyc.com

is called the UNL KB. From an *extensional* point of view, the UNL KB can be viewed as a finite set of tuples of the form:

$$<\text{semantic relation}, uw_1, uw_2>$$

which can be graphically displayed as:

$$uw_1 \text{ —semantic relation}\rightarrow uw_2$$

Given the huge amount of tuples that it may contain, the UNL KB is best viewed from an *intensional* point of view as a first order logical theory composed of a finite set of axioms and inference rules. Most of the axioms state plain semantic relations among UWs, now viewed as atomic formulas:

$$\text{relation}(uw_1, uw_2)$$

Besides atomic formulas, the theory contains complex formulas, like the one stating the transitivity of the "icl" relation:

$$\forall w_1 \forall w_2 \forall w_3 (\text{icl}(w_1, w_2) \wedge \text{icl}(w_2, w_3) \rightarrow \text{icl}(w_1, w_3))$$

We can now turn to the tasks the UNL KB is intended to be used for, and get a clearer picture of its concrete contents according to those tasks. The first task we have mentioned is setting the combinatory possibilities of every UW with respect to the rest of UWs and to the set of conceptual relations included in UNL. For any two UWs w_1, w_2 and any conceptual relation r, the UNL KB should be able to determine whether linking w_1, w_2 with r is allowed (makes sense in principle) or if it is against the intended use of w_1, w_2 and r. If we view the KB as a theory, the question is then if the formula $r(w_1, w_2)$ is a consequence (a theorem) of the set of axioms that form the KB or it is not. The axioms needed for answering such questions are mostly derived from the intended usage of the UNL conceptual relations and the broad semantic classes each UW belongs to.

Thus, the UNL ontology has been developed with several considerations in mind. One of the main characteristics of UNL is its flexibility both formally and linguistically. From a linguistic point of view, the UNL ontology serves to a wide variety of natural languages. From the formal point of view, its integration with other support languages (HTML, XML, OWL) could be easily achieved.

Essentially, UNL has the capability of representing knowledge. However the classical problem emerges. It is the semantic validation process, that is, the set of mechanisms able to deduce coherent domain knowledge from existing one. This is still an open problem, that so far has only attained some partial solutions based on the application of the logic verification rules. However verification rules are not enough to establish a model with sufficient semantic coherence.

REFERENCES

[1] Ballesteros, L. and Croft, W.B., "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval", in Proceedings of ACM SIGIR Conference, 20: 84-91. 1997.

[2] Oard, D., "Alternative Approaches for Cross-Language Text Retrieval". *Proceedings of the AAAI Spring 1997 Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[3] MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web. Proceedings AAAI Spring. Symposium on Cross-Language Text and Speech Retrieval. Menlo Park CA, 1997.

[4] MIETTA project: http://www.mietta.info/

[5] MuchMore project: http://muchmore.dfki.de/

[6] LIQUID project: http://liquid.sema.es/

[7] Gruber. T. A. "A translation Approach to portable ontology specifications". *Knowledge Acquisition*. vol. 5: 199-220, 1993.

[8] MacGregor, R. and Bates R., "The Loom Knowledge Representation Language". Technical Report ISI-RS-87-188, USC Information Sciences Institute, Marina del Rey, CA. 1987.

[9] Farquhar, A.; Fikes, R. and Rice, J. "The Ontolingua Server: a Tool for Collaborative Ontology Construction". *Proceedings of the 10th knowledge acquisition for knowledge-based systems workshop*, Canada. 1996.

[10] Genesereth, M.R.; Fikes, R.E. "Knowledge Interchange Format. Version 3.0. Reference Manual". Computer Science Department. Standford University. California. 1992.

[11] Neches, R; Fikes, R.E.; Finin, T.; Gruber, T.R.; Senator, T. and Swartout, W., "Enabling technology for knowledge sharing". *AI Magazine*, 12(3):36-56, 1991.

[12] Lassila, O and Swick, R.R (1999) Resource Description Framework (RDF) Model and Syntax Specification. W3C recommendation, 1999. www.w3.org/TR/PR-rdf-syntax

[13] Luke, S.; Heflin, J. SHOE 1.01. Proposed specification. 2000. http://www.cs.umd.edu/projects/plus/SHOE/spec.html

[14] Horrocks I.; van Harmelen, F. "Reference description of the DAML+OIL ontology markup language". 2001. http://www.daml.org/2001/03/reference.html

[15] Yergeau, F; Bray, T; Paoli, J; Sperberg-McQueen C.M and Maler E. "Extensible Markup Language (XML) 1.0 (Third Edition)". W3C Recommendation. 2004. http://www.w3.org/TR/REC-xml/

[16] Bechhofer. S; van Harmelen, F.; Hendler, J.; Horrocks, I.; McGuinness, DL; Patel-Schneider, P.F. and Stein, L.I. OWL: Web Ontology Language Reference. W3C Recommendation. 2004. http://www.w3.org/TR/owl-ref/

[17] Quillian M.R. Semantic Memory. Semantic Information Processing. M.Minsky (Ed.), MIT press. 1968

[18] Wood, J. "What's in a link?" In R. Brachman and H. Levesque (ed) *Readings in Knowledge Representation*. Morgan Kaufmann. 1985.

[19] Burg, J.F.M. and van de Riet, R.P. "The impact of linguistics on conceptual models: consistency and understandability". *Data & Knowledge Engineering*, Vol 21, 131 – 146, 1997.

[20] Shamsfard, M. and Abdollahzadeh, A. "Learning Ontologies from Natural Language Texts". *International Journal of Human Computer studies*. Vol. 60, 17-63, 2004.

[21] Bateman, J.A; Henschel, R. and Rinaldi, F. "The Generalized Upper Model 2.0." 1995. http:// http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/gum/index.htm

[22] UNL Center. UNL specifications v 2005. http://www.undl.org/unlsys/unl/unl2005-e2006/