Developing Ontological Foundations for Interlingua

Kavi Mahesh
CRL, NMSU
mahesh@crl.nmsu.edu

1. What is an interlingua?

I would like to reiterate the position that an interlingua must be grounded in an ontology, especially in a multilingual context (i.e., for MT of more than one source or target languages). It is not possible to determine what an interlingua is without considering the purpose(s) it is intended to serve. MT is too nebulous a task; in the following, let us assume that interlingual representations must in fact support both interpretation and transfer:

- extraction of source meanings including word sense disambiguation, disambiguating literal interpretations from metonymic, metaphoric, and other non-literal interpretations, resolving syntactic and semantic ellipsis, coreference, etc.

- transfer of syntax, stylistics, non-literal usage, ambiguity, etc. when possible: e.g., word sense ambiguities can be carried along when the target language provides an equivalent ambiguous word

2. What information is captured by an adequate interlingual representation system?

Consider a related question: "what information must be captured in the ontology to support interlingual representation?"

- a mere taxonomy of primitives (or concepts) is NOT sufficient

- a rich set of inter-concept relationships must be present, especially to support meaning extraction in the presence of ambiguity, non-literal usage, semantic ellipsis, etc.

- default knowledge must be present: selectional constraints on inter-concept relationships must be "tight" to be useful. It is not prohibitively expensive to acquire precise constraints; it is much more practical to acquire constraints at two different levels: a default constraint that is as tight as possible and an overall

constraint that is largely inclusive but still useful.

- uniform coverage of any kind of knowledge is of utmost importance. For example, if default values of attributes are included for a concept, they must be available for all other concepts that could have the same attribute. Without such uniformity of coverage and uniform grain-size of representation, knowledge acquired at great cost turns out useless during processing.

On the other hand, what information is not needed for MT?

- formal definitions are not needed; intuitive descriptions of concepts and their properties are sufficient. Formal definitions (in the form of necessary and sufficient conditions for each concept, for instance) are prohibitively expensive to acquire on a large scale.

- there need not be a well-defined distinction between every pair of siblings: e.g., the real difference between WALK and RUN is not useful for MT.

- moderate granularity and limited expressiveness of interlingual representations are indispensable virtues in practice. Almost any meaning can be decomposed into arbitrarily detailed and complex representations; we must limit this tendency and live with a coarse interlingual and ontological representation for practical reasons.

3. How can interlingual representation systems be built or scaled up?

Ontological (and lexical) knowledge is best acquired by following a situated development methodology: that is, every piece of knowledge acquired must be required for solving a real problem in a real MT situation and it must be put to use and tested immediately upon acquisition. Close cooperation among lexicographers, ontologists, domain experts, MT system developers and testing teams is inevitable for successful knowledge acquisition. An ideal situation is one where the ontology and lexicons (for both analysis and generation) for at least two different languages are being acquired simultaneously.

From our experience in building the Mikrokosmos ontology, we can claim that:

-10-20 person years of effort is sufficient to acquire a sufficiently broad ontology (about 50,000 concepts) with sufficiently rich inter-concept relationships and constraints

- the cost of acquiring such an ontology is NOT significantly greater

than the cost of acquiring a lexicon with sufficiently rich semantic information, i.e., it does not introduce an unsurmountable bottleneck any more than what we already have in lexicon acquisition for interlingual MT

- ontologies are much more reproducible than many people think. There are striking similarities in concept organization and classification across all major ontologies (Cyc, Mikrokosmos, Wordnet, Sensus, etc.). It is not unthinkable to agree upon a common ontology for MT or merge previously acquired ontologies to build a broader foundation for interlingual MT.