****

First Steps toward Building Interlinguas of Scale

Eduard Hovy
USC Information Sciences Institute
hovy@isi.edu
August 1996

In the construction of an Interlingua for Machine Translation a system, two principal challenges stand out:

1. the design of a representation approach simple enough to be manageable by human representers, yet sophisticated enough to be able to capture meaning in a truly interlingual manner;

2. the construction of an Interlingua 'lexicon', a set of representation terms enough to provide adequate coverage over a real-world domain yet consistent enough internally to be manipulated by automated processes.

The former challenge is the domain of ongoing efforts in lexicography and knowledge representation [Nirenburg et al., Copestake, Dorr], among others. It is a complex endeavor with little assistance from automated procedures — the crucial design work must be done by humans, and the development of methodologies, rigorous performance procedures, and criteria of evaluation is ongoing.

The latter challenge is the focus of the proposed presentation. I will (if accepted) discuss the creation and testing of an Interlingua 'lexicon' of large enough scale to support open-domain translation (as well as of various other NLP tasks). It may seems strange—some would say impossible —to separate the creation of an IL 'lexicon' from the actual representation. However, by IL 'lexicon' I mean here a termbank, a set of symbols containing only a very sparse semantics, that serves as a pivot structure between the lexica of the various languages. I am not talking about an Interlingua Lexicon in the full, rich, sense.

Again, some may question the purpose of constructing such a set of 'empty' symbols. The answer is simple: practical experience has shown that it is possible to achieve wide-domain MT (and a variety of other NLP tasks such as IR) with such symbols. Such a set provides a baseline for IL-based MT performance. Admittedly, the quality is not always very good; that is the reason for continuing work on enriching the symbols' contents. But proceeding "from the outside in" (large set of symbols, gradually enriched), as opposed to "from the inside out" (small set of richly annotated symbols, gradually grown in number), is an important approach, for it tells us several things:

1. roughly, the semantic 'regions' in which symbols are required (for example, the 'region' of speech act representation, or the region of attributes);
2. a sense of the levels of representational delicacy required in each such 'region' relative to other 'regions' (measured by, say, the approximate numbers of such symbols relative to the numbers in other 'regions');
3. an indication of the organization of such symbols and the underlying interrelationships that prove most useful (for example, ISA vs. PARTOF vs. SYNONYM);
4. an overall framework of anchor points into which we can embed more delicately articulated symbol sets, whenever they are forthcoming;
5. a way of helping to bring different, competing, enriched symbol sets (or domain theories) into correspondence and to compare them.

Some of these ideas may be controversial, and I will not belabor them. Instead I will describe some work I have recently performed in service of creating a new, large-scale, IL termbank, organized as a property inheritance taxonomy. The goal of this work is to establish a kind of 'standard' Reference Ontology (R.O.) which would be available to anyone via the Web, for the following uses (among others):

- to allow interprocess communication for systems developed at different sites (when both systems' terms are translated into the IL terms),
- to allow the comparison of different domain models for the same domain,
- to identify areas of shortcoming in the R.O. and invite suggestions on how to fill them.

This effort is a collaboration of the following researchers:
- IBM San Jose (Bob Spillers, Andras Kornai) (lead organization),
- USC/ISI (Eduard Hovy),
- CYCorp (Doug Lenat, Fritz Lehmann),
- Conceptual Graphs (John Sowa),
- Stanford University (Bob Engelmore, Adam Farquhar)

and more. It is part of an ANSI Standards group working on representations.

The work I will describe represents the first steps in creating an R.O. These steps bring together the uppermost regions of several large ontologies and relate their terms to one another, to the extent this can be done. The result is a taxonomy, viewable either as a single integration of the terms from each of the constituent ontologies, or as each constituent alone, with pointers to the others as appropriate.

The constituent ontologies (with, parenthesized, the number of concepts under current consideration) are:
1. The Pangloss Ontology Base from ISI (approx. 300 concepts)
2. The 'top' concepts from CYC (approx. 1500)
3. The top concepts from EDR (approx. 100)

The result is represented in SENSUS, a simple KR system analogous to ART, KEE, Loom, FrameKit, and the like.

How one goes about integrating (or at least, finding correspondences and linking together) such disparate symbol sets is not exactly clear. I will describe the six-step process I followed, providing in a handout the inputs and results of each stage, and run some of the transformation programs as part of the discussion. I will highlight the easy and the difficult aspects of the task, and point out some problems caused by the idiosyncrasies of the various symbol sets.

If there is a discussion period, I would be extremely interested in hearing opinions as to how this work should continue, what results (if any) would be useful to researchers, and (if useful) in what form.

****