****

Interlinguas: natural languages, logics or arbitrary notations?

Yorick Wilks
University of Sheffield
yorick@dcs.shef.ac.uk

What are interlinguas, and does the answer have any practical effect on the usefulness, success, or otherwise of interlingual MT, a paradigm that still has life left in it, and some practical successes, certainly in Japan. In deference to the gathering, I will focus what I have to stay on the interlingual representation language, as you might expect to find it in an MT system—normally some pidginised version of English with a counterintuitive syntax—but what I have to say applies more generally to symbolic knowledge representations, including those applied to MT (e.g. KBMT) and those in the mainstream AI tradition.

If we take the view that they can be arbitrary notations then many more systems come within the class of interlingual machine translation systems than would normally be thought the case: certainly SYSTRAN in its earlier periods (before the point at which it was declared a transfer system) when the results of a source language analysis were stored in a complex system of register codes and, most importantly, this was done for more than one source language—thus giving the storage codings, which were largely arbitrary in conventional linguistic terms and certainly without comprehensible/readable predicates, a degree of linguistic "neutrality" that is thought to be part of what is meant by an interlingua.

Taken more strictly, SYSTRAN was never an interlingual system because its power came largely from its bilingual dictionaries, and, as a matter of definition, a bilingual dictionary is a language-pair dependent transfer device. At that level of strictness, there have been very few true interlingual MT systems, (i.e. without a bilingual dictionary). Probably the only candidate is Schank's MARGIE system of about 1972, which did some English-German MT via a conceptual dependency (CD) representation. Schank's CD representation is also much closer to the stereotype of an interlingual representation, having a range of English-like predicates (TRANS being the best remembered) within a specified syntax and diagrammatic notation.

My own small English-French MT system, contemporary with Schank's and

in the same CS department was therefore not interlingual, even though our representations had much in common, because I believed a bilingual dictionary essential, and that the interlingual representation and its associated algorithms selected the correct equivalent from among a set of candidates the lexicon provided. Schank believed then, and I denied, that any such crypto-transfer mapping was needed, but only he Source-to-interlingua and interlingua-to-target translators. It was Charniak who later supplied argument's arguments whose force I already felt, that no level of coding at such a grain could be expected to distinguish in output: sweat, sneeze, dribble, spit, perspire, as well as a range of less attractive possibilities all associated with the Schankian primitive EXPEL taken along with a coding for LIQUID.

The issue of grain here was obviously a function of the richness of the interlingual vocabulary (Schank then had about 14 primitive actions and I about 100 primitives of different syntactic types). IF the interlingua had the resources of any natural language, then those distinctions could have been made and that of course focuses exactly the question of what it would mean for an interlingua to have the resources of a natural language as opposed to being a formal language with primitives that may appear to be language-like, as Schank's certainly did, but which their author's deny are in fact language items, let alone English words.

That position of denial is not to be found only in the 1970's: Schank's position is essentially that of Nirenburg and Raskin (1996), when supporting Mikrokosmos as a "language-neutral body of knowledge about the world" in contrast to the "recurring trend in the writings of scholars in the AI tradition..toward erasing the boundaries between ontologies and taxonomies of natural language concepts".

This dispute can take unhelpful forms such as "Your codings look like natural language to me"; "No they dont". Moreover, it is not clear that any settlement of this issue, for either side, would have effect whatever on the performance or plausibility of interlingual MT systems. One can can also filter out more full blooded versions of the NL-IL identity, such as the Dutch MT system based on Esperanto (an NL for the purposes of being dismissed from this argument) and the reported systems based on the S. American language Aymara, said to be without lexical ambiguity (see below) and therefore ideal for this role. These cases, whether or not they work in practice, fall under the criticism of Bar Hillel in the earliest discussions of interlingual MT that having an NL in the IL role would simply double the work to no benefit by substituting two MT tasks for one.

The interesting issue, given that we can all concede that ILs do not

normally like quite like NLs, and do certainly have some superficial features of formal languages (brackets, capitalization, some non-linguistic connectives etc.) is whether they have any significant features of NLs, which I would take to mean above and beyond the simple appearance of a number of NL words in their formulas, normally drawn from English. English words appear in profusion in many programs particularly those that encourage arbitrary predicate naming, such as LISP and Prolog, from which one could not of course conclude that programs in those languages were IN ENGLISH. Appearance of words is not enough, or French could be declared English, or vice versa, or Roumanian Turkish.

The feature I would seize on in discussion is whether the primitives of interlingual formalisms suffer ambiguity and progressive extension of sense as all known NLs do (except perhaps Aymara, but that may reflect lack of study). Some formal languages can tolerate substantial ambiguity of symbols—early LISP, which functioned perfectly well, is now conventionally said to have had a key symbol (NIL) three-ways ambiguous. LISP programs presumably implicitly resolved this ambiguity in context, or did they?

It is widely believed that NLs have their ambiguities resolved in use, up to some acceptable level, and that extensions of sense take place all the time, whether rule governed (e.g. as in Pustejovsky's generative lexicon, deriving from Givon's early work in the 1960s) or, as in the old AI/NLP tradition by means of manipulations on lexicons and knowledge structures that were general procedures but not necessarily equivalent to lexical rules. It would it be like, and I have no clear answer, to determine that the primitives of an IL representation were in this position, too. Schank did, after all split the early TRANS into MTRANS, ATRANS and then others, so the suggestion has precedent.

An answer might require a trip through the more empirical aspects of recent NLP/CL and ask what evidence we have that any given symbol in its occurrences in corpora has more than one sense? Has this any empirical, non-circular answer, that does not require appeal to existing polysemous dictionaries or contrastive translations? I believe the answer is yes, and that an IL does have this key feature of an NL and that no disastrous consequence follows from thus viewing ILs as reduced NLs, rather than full ones. This has for me more intuitive plausibility than continuing to maintain that what seem to be NL features of ILs are not in fact but are language independent. That claim always seems to me one that it is impossible to defend in detail but is a clear residue of the hypnotic power of

intuition in an age of empiricism and calculation.

****