****

Position on Interlingua:
John White
PRC
white_john@po.gis.prc.com

What really ought to represented in the intermediate representation?

I suspect that few people dispute the potential benefit of a language-independent representation in an MT model, and the interlingua that expresses the representation. Imagining what kinds of problems would have to be addressed there, it is easy to find tempting candidates for interlingual representation that create real problems for accepting the principle in the first place. Many of these problems live in the metaphysical, the knowledge-based, pragmatic, sociolinguistic, etc. trappings of language in use, and the possible different expectations of a native speaker of source and target. But there are some problems that have implications for plain old morphological or lexical forms, which give one pause. These problems have to do with how much contextual semantics work you must do in analysis (and represent in the IR) to cover really language-specific demands.

Here are two such problems, one with an apparently easy fix, which may suggest a way to fix the harder one.

The easy one:

Chinese dialects, Mayan languages, and, apparently Navajo, share a means of representing the lexical class of nouns by expressing a numeral classifier. These are like the English "round" (of ammunition) and "head" (of cattle), except that they are more fully specified over most or all of the count nouns in the language. Semantics are generally attributed to the selection of numeral classifiers; thus Tojolabal-Maya "wan" seems to express humanness, and may appear in expressions like "chab' wane? winik (two person-units man)", "osh wane? ishuk (three person-units woman)", and so on.

Given the apparent the observation that the classifier can be predicted by lexical semantics, and the fact that the phenomenon occurs in many languages, it is tempting to try to "handle" this somehow in the intermediate representation, by expressing the semantics for it in the interlingua. If the target language uses numeral classifiers, generation would make classifier selection based on that semantic representation.

What inserts some reality into the idealistic desire to represent the phenomenon in the interlingua is the fact that the numeral classifier languages do not organize the lexical universe in the same way. That is, the fact that Tojolabal-Mayan classifies "sky" as something wide and thin ("hun lame? sat k'inal" — one slice sky) does not in anyway predict that the other numeral classifier languages do. This means that in order to handle the phenomenon in the intermediate representation, the IR would have to express all of the semantic

properties known to be sensitive in any classifier language, and hope that there are no contradictions (which there will of course be).

This problem is easy because the practical solution is also the right one, namely to generate the numerical classifier solely from the language specific requirements on the target side. Each lexical item will have some sort of specification about the classifier it expects. In the case where a noun may take different classifiers in different instances, the lexical description on the target side will contain variables instantiated from the normal compositional semantics of the intermediate representation.

Here is the hard problem:

English generally distinguishes between flesh on the hoof and meat on the plate (pig-pork, sheep-mutton, cow-beef, etc.). Spanish and German generally do not. We know from this that the intermediate representation must have done some work, possibly difficult work, in inferring from context whether a reference to animals or parts thereof is a reference to flesh or meat, so that the interlingua can express sufficient semantics for English generation.

If we do this, then we can rest assured that regardless of the source language, we will be able to generate the correct distinction in English. But what if I am translating from Spanish to German? My intermediate representation has done all of the inference work to generate the flesh-meat distinction, but in this pair I don't need it. Why should an MT system do substantial work that the pair doesn't need?

What makes this question hard is the fact that the English-required distinction can't be done on the target lexicon side, as it can for numeral classifiers. A great deal must be known about whether something is intended to be eaten, and what is doing the eating, etc., information that has to come from all over the place in the source language expression. So it appears that this distinction must indeed be handled in the IR.

But doesn't this mean that the IR has to express the semantics of all the lexical idiosyncrasies of all languages? Surely every language has some lexical phenomenon similar to English flesh-meat. Doesn't this leave us in the same trap from which we escaped in the numeral classifier case?

I don't know the answer to this particular problem. But I think the solution has to do with the level of generality at which we do our

interlingual representation. It may well be that a great deal of the semantic work specific to a language must be done at generation time, possibly even after a round of lexical selection. In this model, a reasoning tool examines a partially or fully lexicalized target representation, and makes a judgment about its felicity (semantic, pragmatic, discourse-wise), choosing alternates in some cases and lexicalization of variables in others. This delegation of powerful reasoning to the generation component seems to violate our current sensibilities about the role of the interlingua, but the interlingua model of MT remains language independent, and in fact becomes more so by expressing only what is truly universal and not by trying to be all things to all languages.

****