****

The 'Lingua' in Interlingua

AMTA SIG-IL Workshop, 1996

Robert Belvin, Bonnie Glover Stalls, Christine Montgomery, Alfredo Arnaiz
Language Systems, Inc.
robin@lsi.com

An interlingua can be defined as a metalinguistic representation of the function of a linguistic object which is not dependent on the language-specific form of that object. In Language Systems, Inc. (LSI)'s multilingual Machine-Aided Voice Translation (MAVT) system, the interlingual representation consists of a set of event and object frames with slots that are filled with information derived from or associated with the text that is being processed. These slot fillers include information bearing on propositional content as well as communicative intent, pragmatics, and other kinds of information that are present to varying degrees of explicitness in the text. They also provide a means by which contextual and domain knowledge that has no realization or is ambiguous in the text can be used during the translation process. To a great extent the information filling these slots is not language-specific; however, there are some interesting ways in which some language-specificity is preserved and not only does not interfere in the translation process but actually facilitates it.

As David Farwell (ACL 1994) has pointed out, the goal of an interlingual representation is not "language-independence" but rather "language-neutrality". We essentially agree with this position, but would further suggest that it is not necessary to strip away from the textual representation all vestige of the source language, but rather to render it in a neutralized form that is easily mappable into any potential target language. An interlingua for an MT system, which must be capable of transmitting information from one language to another, is a kind of language, or representation of language, for which the requirement for neutrality need not limit its expressive power. In fact, as discussed at some length in Dorr (1993), preserving certain aspects of the linguistic structure of the text can help to minimize the need for a deeper level of conceptual representation and to construct the target language text. A case in point is the incorporation of lexical conceptual structure (LCS) into the interlingua in LSI's MAVT system as a means of ensuring that a verb or other predicate with an appropriate predicate-argument structure is selected and appropriately saturated in the target language.

Our position has been that predicate-argument relations are more efficiently represented AS predicate-argument structures than other kinds of representations. That is, since the information encoded in a P-A structure readily expresses the relations of interest, it seems unwise to transform that representation into some other type of structure. This seems especially true in employing an interlingua in an MT system. Our experience has led us to the belief that the use of a quasi-syntactic representation of eventuality (i.e. event or state) concepts has facilitated the translation process, especially w.r.t. lexical selection of target-language predicates, and generation of target-language syntactic structures. If one regards the creation of an interlingual representation as a series of steps which undo the language-specific packaging of the relevant information, the complete undoing of the syntactic organization of eventuality representations appears to be a step which may not only be unnecessary, but undesirable.

For example, when matching a concept in an interlingua to a lexical item in the generation stage of translation, the use of quasi-syntactic lexical-conceptual structures has allowed us to collapse a process which would otherwise require two steps into one. The characteristics of our LCSs usually allow us not only to find a lexical item (or items) in the target lexicon which matches the relevant concept (both in its core meaning and selectional restrictions), but we can simultaneously check if the candidate lexical item has an appropriate subcategorization frame.

It should be mentioned that we arrived at the decision to employ semantic structures modeled on Jackendoff's Lexical-Conceptual Structures in our MT system after considering several alternatives. Jackendoff type LCSs appeared more desirable than the alternatives because (i) they provided a way of representing concepts which partially solved word-sense disambiguation problems without relying on language-specific predicates and (ii) they facilitated mapping between interlingual concept representations and language-specific syntactic representations. This latter advantage seems to be due to the fact that they are structured in much the same way as a syntactic predicate. In addition, since our predicate concepts are quasi-syntactic, the possibility arises that the same kind of constraints can be applied to them as are known to apply to syntactic representations. This kind of parallelism between lexical-conceptual structure and syntactic structure has in fact been argued to reflect a genuine psychological reality based on various types of linguistic phenomena in research by Hale and Keyser (1993), Bouchard (1995), Jackendoff (1993), and others. While we are not necessarily committed to this premise, the possibility nonetheless provides additional support for the validity of the approach.

Our work in developing an interlingual MT system grew out of our Data Base Generation (DBG) system, which was developed over a number of years and which analyzed text and produced output for a variety of downstream applications, including information extraction and retrieval, and message fusion. The goal of the DBG system was to instantiate a set of event and object frames, called "templates", which represented the content of the text being processed. At the topmost level were meta-templates, which represented the meta-event of writing and sending the text being processed, and so could incorporate higher-level discourse features of the text (e.g., source of the text, time it was written, recipient, and so on).

One thing that we discovered in working with a variety of applications is that the content of an adequate representation "depends on the application." What is adequate for one type of application may be completely inadequate for another. For example, in highly regimented contexts such as written communications reporting flight activity of aircraft by military surveillance teams, there is virtually no need for anything but the scantiest information on communicative intent, since the communicative intent remains constant throughout the reports. Other factors, such as the degree of belief of the writer in the facts being reported, however, are highly significant and must be analyzed and represented. In a voice translation system designed for interrogation purposes, identifying communicative intent in the source speech and providing a reasonable approximation in the target speech is very important because the intent is highly variable, and the response of the hearer may be very sensitive to it; it must therefore be given some representation in an interlingual representation.

An interlingua is a kind of knowledge representation (KR), very similar in many ways to the KRs that we have worked with previously. One characteristic of an interlingual MT system such as the one we have developed which distinguishes it from many other interlingual MT systems is that it organizes concepts according to two different taxonomic schemes. One is a lexically oriented conceptual scheme (the LCS taxonomy), the other a typical (non-lexically oriented) KR scheme. In MT, using the interlingua as a means of preserving the structure of the source language sentence for use as a kind of filter or guide in selecting target language lexical items and syntactic structures, makes a good deal of sense. Conceptual information is certainly an inherent part of the process, and in the MAVT system is available when needed. Nouns in the lexicon are indexed to nodes in the conceptual hierarchy, and selection of the target language nouns is done by selecting the nouns associated with the same or related nodes (ontological entries) as those in the source language. However, for verbs the relations among verb concepts in the concept hierarchy are used primarily in cases where there is no exact LCS match. In that case, adjacent nodes are checked for possible near-matches that can incorporate the information in the interlingual representation.

This dual taxonomy strategy allows to take advantage of the virtues of lexically structured concepts where possible, but allows us to exploit non-lexically structured concepts when necessary. This organization, as we have suggested, is desirable in an interlingual MT application, but may be unnecessary in other types of applications. This is because the most obvious virtue of lexically structured concepts is that they facilitate target language generation. In a text understanding application, it may be preferable to bypass lexically structured concepts and map directly to a non-lexically structured knowledge base.

REFERENCES

Bouchard, D. (1995) The Semantics of Syntax, U. of Chicago Press, Chicago.

Dorr, B. J. (1993) Machine Translation: A view from the lexicon, MIT Press, Cambridge.

Hale, K. and S. J. Keyser (1993) "On Argument Structure and the Lexical Expression of Syntactic Relations," in Hale and Keyser, eds. The View from Building 20, MIT Press, Cambridge, pp. 53-109.

Jackendoff, R. (1993) "X-bar Semantics," in Pustejovsky, ed. Semantics and the Lexicon, Kluwer, Dordrecht, pp. 15-26.

****