PREceedings
of the
Pre-Workshop on ILs and IL Approaches to MT
[1 October 1996, at Second Conference of the
Association for Machine Translation in the Americas, Montreal, Quebec, Canada]

****

Stephen Helmreich
Computing Research Laboratory
New Mexico State University
shelmrei@crl.nmsu.edu

WHAT IS AN INTERLINGUA AND WHAT INFORMATION SHOULD IT CONTAIN?

I. INTRODUCTION

The traditional defining characteristic of an Interlingua, according to Hutchins & Somers (p. 73), is that it "is neutral between two or more languages" so that, in principle, given a representation of an utterance in the Interlingua, the source language of the utterance cannot be determined from that representation. From this follow the other general features of an Interlingual MT system: the independence of analysis and generation, the use of language-independent knowledge sources, the attempt to represent the "meaning" of the text using the Interlingua, the claim to "universality," and the abstract nature of Interlingual representations.

In this position paper, I argue against this characterization and suggest that a better characterization is that an Interlingual representation is one geared to explicitly represent the intent of the text author(s). Consequences of this position for the information contained in an interlingua are also somewhat fleshed out.

II. IT IS IMPOSSIBLE TO ACHIEVE HIGH-QUALITY, AUTOMATED MACHINE TRANSLATION WITH A "LANGUAGE NEUTRAL" REPRESENTATION AS THE SOLE INPUT TO THE TARGET LANGUAGE GENERATOR.

The conclusion stated immediately above follows directly from Bar-Hillel's argument (1960) that, in principle, any bit of knowledge might be required in order to disambiguate some text. While this argument is usually made in support of the need for knowledge-based interlingual MT, it can also be used _against_ an Interlingual approach as characterized above. For surely, one of the possibly significant bits of information that might be necessary to adequately disambiguate a text is precisely the original language of the source text and the actual source language text itself. Thus, this information must be part of any Interlingual representation of the text, if an adequate target language text is to be generated from it. But then, of course, it is not interlingual in the sense described above, that is, containing no hint about the language of the original text.

A simple example might help. A cartoon sequence in the local paper was, for a while, printed with only Spanish text. (After protests from English-speaking readers it was then printed with English sub-titles as well, and then later dropped.) One sequence included the following dialogue. In this case the visual element is non-essential, but the background was a park filled with playground equipment.

?Por qui el pollo atraverss el parque?
!Para ir al otro tobogan!

This apparently inane dialogue makes sense if one hypothesizes an original English dialogue as follows:

Why did the chicken cross the park?
To get to the other slide!

Leaving aside for the moment the cultural specificity of the joke-type itself, given this reconstruction it is clear that even to begin to translate this dialogue appropriately, it is necessary to note the phonological closeness (rhyme) between the two English words "side" and "slide." Thus, this information must be available in the interlingual representation. But then it contains a reference to the source language itself.

It cannot even be plausibly argued that this information could be extracted and represented in the Interlingua in some non-language specific format. The aspects of the source language text that might be relevant are, in principle, unlimited, and could include information about the phonetics, phonology, morphology, syntax, or orthography of the source text.

In conclusion, to ensure high-quality translation at all times it is necessary to include language-specific information about the source language text in any representation that will serve as the generation input.

III. INTERLINGUA AS A MEANS OF REPRESENTING THE INTENTION OF THE AUTHOR(S).

Let me make it clear the claim in II. above does not imply support for a transfer-based approach to MT. Far from it. Indeed, in the example above, no transfer-based approach could begin to succeed. Critical to finding an appropriate translation for the dialogue is the recognition of the author's intent to tell a joke, the recognition of the genre of joke-telling (i.e., having a punchline) and then recognizing the subverting of that genre by having a non-punchline punchline as the key to the joke. In addition, it involves recognition of a particular joke as the underlying form on which this is a witty(?) takeoff.

It seems to me that the difference between an Interlingual approach and others is not on what the IL representation cannot contain (i.e., any reference to the source language), but rather on the depth of analysis contained in that representation.

And, as in the above example, it is not sufficient even to represent the "meaning" of the language involved. In order to provide an adequate translation, the intent of the author(s) must be ascertained. Therefore, I suggest, a better characterization of an Interlingua is as a representation of the source text which allows for an explicit representation the intent of the author(s). This requires analysis to a depth greater than that which even current IL/KB MT systems, such as Mikrokosmos, perform. It requires reasoning from the semantic propositions expressed plus relevant information in the Knowledge Bases in the system (plus, on occasion, information about the actual sounds, words, and structure used to express those propositions) to the intent of the author(s) of the text.

IV. CONSEQUENCES OF THIS POSITION.

Changing the characterization of an Interlingua from a formal categorization to a categorization based on content affects none of the the following characteristics of an Interlingua: knowledge-based processing, representation that includes the "meaning" of the text, universality or abstractness of representation. It should not even affect the independence of analysis and generation in that there do not need to be special rules in the generation procedure that specifically are cued to the source language (though they could not be ruled out by this definition).

In addition, the following features would seem to be, if not required, at least consonant with this definition.

A. Explicit representation of the beliefs of the author(s) and of other participants in the communicative event.

B. Explicit representation of various levels of author intent, e.g., locutionary, illocutionary, and perlocutionary intent.

C. Inferencing/control structure that is non-monotonic. Since the intent can only be deduced, not observed, often on the basis of default or other types of non-deductive reasoning, the intent can be determined only with a certain level of confidence.

D. Emphasis on the establishment of coherence of interpretation.

E. Explicit representation of the chains of inference used to determine the author's intent.

V. CONCLUSION.

It has been argued that an interlingua which cannot represent aspects of the actual source language text is insufficient as a basis for high-quality machine translation. Indeed, as the intent of the author(s) become more complex and their attention to the details of the text more conscious (as in poetry or other emotive uses of language), more and more of the physical aspects of the text will be vital clues to achieving even an adequate translation. It is suggested that an Interlingua be defined as one geared toward to the representation of all aspects of the author's intent.