[Presented at the Conference on Mechanical Translation
at Massachusetts Institute of Technology, June 1952. Not published]

The Treatment of "idioms" by a Translating Machine

Y. Bar-Hillel

One of the standard objections I heard often raised against the possibility of MT was its alleged inability to cope with idioms. I shall try to show that machines can be constructed that would know how to treat idioms; the major problem is not to find a method for this treatment but to find a reasonably efficient combination of the various methods available for this purpose.

Let me first state clearly what an "idiom" would mean for a machine: Whenever none of the sentences of the target-language TL corresponding to a given *s*entence of the source-language SL, according to a certain set of grammatical rules and a certain dictionary, is regarded by an appropriate authority as a satisfactory translation of the original sentence, I say that this sentence is *idiomatic* with respect to this TL, the set of rules, the dictionary (and for this authority). For our case: Whenever the post-editor will be able to pick out of the many correlates to the original sentence offered by the machine even a single one that may be regarded as a satisfactory translation, the original sentence is idiomatic. If, for instance, the dictionary presents as the correlates of the words occurring in the German sentence.

Es gibt einen Unterschied

it (he, she), *gives*, *a (one), difference*, respectively, if the translation rules state that a sentence of this form may be translated into English word-by-word, and if none of the resulting English sentences is regarded as a satisfactory translation, then *Es gibt einen Unterschied* is idiomatic (with respect ...). Notice that, for the time being, it makes no sense to put the responsibility for the idiomatic behaviour of this sentence on its initial phrase *Es gibt*, though this may perhaps be done at a later stage.

How, now, could a translation machine overcome these idioms? The answer is obvious: Not at all. And this simply by our definition of the idioms. The only way for a machine to treat idiom successfully is – not to have idioms! And this, of course, is the responsibility of the human being who sets up the grammatical rules and the dictionary. Our problem is, therefore, not how should a machine treat idioms but how could one make sure that the machine will not be confronted with idiomatic expressions.

Now, this can be done in various ways. I shall outline two or three of the methods which, each individually, are theoretically self-sufficient. I shall not do this in abstracto but only with respect to the German sentence mentioned before. It should then however be sufficiently clear how these methods would work in general.

Let us therefore consider again the sentence

Es gibt einen Unterschied.

The first method consists in supplementing the standard dictionaries in such a way that the satisfactory English translation

There is a difference

will be forthcoming as one of the possible correlates. This can be achieved, for instance, by having *there* as an additional correlate of *es*, and *is* as an additional correlate of *gibt*. As simple as that.

This method probably sounds as preposterous to you as IT sounds to me. But why does it? I am not sure what you would answer somebody who insists that this is the thing to do. "It works," our obstinate reformer might argue, "and no damage is done, since none of the former legitimate translations is overthrown". Well, the only reasonable answer I can see is that it works too well. In addition to the welcome combination *there is*, many other gratuitous combinations will be introduced, the elimination of which through consideration of context might beat least troublesome, sometimes perhaps impossible. To have to cope with *she is a doll* as one of the possible translations of *sie gibt eine Puppe*, even if this translation would be

excluded through the context in which the German sentence is embedded is certainly too high a price to pay for the elimination of one idiomatic expression. If I am not mistaken, *es gibt* is the only phrase that might encourage us to have *is* as a correlate of *gibt*. Were there more of such phrases, say a hundred, then it would probably be worth-while to have *is* as a correlate. Where to draw the line is a question of expedience which I am, of course, in no position to answer. In practice, one would like to have another method to deal with cases as the one considered here.

A promising method is the following, which consists essentially in a change of the standard grammatical rules. Just supplement the ordinary word- or stem-dictionary by a special phrase dictionary whose entries will be exactly those phrases, a word-by-word translation of which would turn out to be unsatisfactory. For our case, the phrase dictionary would contain *es gibt* as one of its entries with *there is (are)* as the correlates of this entry. Notice that sometimes certain grammatical rules will have to be applied before the phrase dictionary will be invoked. One such rule will have to deal with the translation of question sentences like *gibt es einen Unterschied*? Notice also that the fact that *es gibt* would appear in the phrase dictionary does by no means imply that ***all*** tokens of this phrase will have to be translated by *there is (are).* In general, this will only be an ***additional*** possible translation. *es gibt* in *es* (*das Mädchen*) *gibt mir einen Kuss* will certainly not be rendered by *there is*. In some cases, however, the so-called literal translation may never be to the point. The instruction for the machine (as well as for the dull student) will be to hunt always first for the possible occurrence of idioms in the given sentence and to indicate in the phrase dictionary whether the correlate to some phrase is the only possible translation or whether "literal" translations should also be considered.

This second method is, of course, theoretically completely foolproof. The only practical drawback is the size of the resulting phrase dictionary. Again, I do not know how many entries we can afford to have in the phrase dictionary. It would certainly be very unwise to have in the regular English-German dictionary for the entry *fair* only, say, *schön* and *nett* as correlates, so that *fair play* would have to be treated as an idiom and would appear as such in the companion phrase dictionary. This is because *fair play* is not the only combination where *fair* cannot be satisfactorily rendered by either *schön* or *nett*.

As a matter of fact, this second method of dealing with idioms is quite customary in many large-scale dictionaries. For machine translation, certain changes in arrangement would be indicated.

The third method is nothing else but a variant of the second one. This variant shows, however, enough interesting features of its own to deserve special treatment. According to this method, no changes would be introduced into the standard dictionaries, nor would a special phrase dictionary have to be compiled. Instead, the reader of the translation would be told that certain target language phrases would, or perhaps only might, be replaced by other phrases. The raw translation of *es gibt* would still be *it (he, she) gives*, but the English reader or, preferably, the English post-editor, would be instructed to replace, or at least to consider a possible placement of, *it (he,she) gives* by *there is (are).*

The main difference of this method as against the second one is, of course, the fact that according to the third method, elimination of idioms is handled on a unilingual basis, in the mentioned example in English exclusively.

I think that the task of finishing a good combination of the mechanica1 methods (and perhaps others), either for human or machine translation, should prove to be interesting not only for the [translator?][1] but also for the theoretical linguist.

[1] Blank space in original typescript [Editor]

Discussion

Locke: The "idioms" Bar-Hillel is talking about are interlingual idioms. I wonder whether there are also idioms for the native speaker of one language, without regard to possible translations into other languages.
Reifler: Yes, there are idioms within one language.
Dostert: I submit that a man who never learned a foreign language would never know what an idiom is.
Oswald: There are all kinds of idioms. I would like to know what the common characteristic of all these various types of idiom is.
Locke: An idiom is something you cannot translate into another language and make sense.
Bar-Hillel: Precisely. An expression of one language is an idiom *with respect to another language* if none of the so-called "literal" translations renders satisfactorily the original sense. I would say that an expression is an idiom within a language, if none of the word-by-word "interpretations", i.e. transformations according to some type of synonym-dictionary, would be regarded as satisfactory. I think that this might do as a first approximation.
Bull: I am fairly certain that one thousand idioms would take care of 90% of the translations that would remain unsatisfactory on a word-by-word basis. If my guess were correct, the whole problem of idioms would be enormously reduced.
Oswald: I think we are back at an old problem. Just as we need an idio-word-dictionary so we need an idio-phrase-dictionary. The more complicated and technical a discourse becomes, the more primitive the language will get, and the fewer the number of idioms that will be used. For such type of discourse, the problems we are fighting to avoid will not be there at all.
Bar-Hillel: The whole problem for me is only one of quantity. How many idioms can you afford to have in a special phrase dictionary without slowing thereby down the translation process in an intolerable degree. Otherwise the best thing would be to have simply all sentences in a sentence-dictionary. The practical problem is to find an upper limit for the number of idioms which can still be handled. The larger the number of idioms, the smaller the number of word-correlates, hence the smaller the load on the post-editor.
Reifler: There will be translations which will not make sense by themselves but will make good sense, in fact render the original meaning, when considered in their whole context.
Bull: Ten thousand idioms would give you 98% of all idiom occurrences even if you translate Shakespeare or Gothic into Chinese. After that you come to a point where an idiom occurs only once in a thousand ages. You can afford to stand this risk.
Bar-Hillel: The main danger is not having sufficient idioms in a phrase-dictionary is not in the fact that some literal translations would be jibberish, it lies in the fact that some of these translations will make sense but the wrong sense and the post-editor will be unable to find this out.
Bull: But we can stand those errors if their number will not exceed, say l%.
Oswald: I despair of ever being able to translate mechanically diplomatic language because I can't read it even in English. One would need a scanner to go in between the lines to get out the meaning of a message.
Dostert: The diplomats say the same thing about linguists. I think that if we are going to restrict MT from the beginning only to high-quality scientific discourse to be edited by eminent scientists, we will wind up by not doing anything at all. To restrict oneself to scientific language may be well and good as a beginning but we have to overcome this restriction as quickly as possible.
Oswald: There is only one thing wrong in your statement. Even if we confine ourselves to scientific discourse, the appeal will be enormous. Also, the scientists have all the money.
Bull: In England, you have no more than 5000 natural scientists out of a population of approximately 50 million. But there are government agencies and other people that are interested in scientific translations. I still think that we should start with scientific translations. As soon as we can handle this, lets go on.
Bar-Hillel: I would like to stress again that this is only a question of quantity. In literary language we shall simply find a much larger number of idioms than in scientific discourse.
Wiesner: I would like to ask a question with regard to human translation. Do human translators work equally in both directions?
Dostert: This is not to be encouraged. It is better to train people to know two or three foreign languages to such a degree that they are able to translate from them into their own native language.
Reifler: During my stay in China, I had often to translate from English into Chinese; I wound up with talking English to Chinese and Chinese to Englishmen, without realizing it.

Dostert: Would it to possible to have a general machine translator in which, through human intervention, a given FL material could be assigned to a certain area of this machine for translation into some specific TL?
Bar-Hillel: This problem of multiple translation as well as that of two-way translation has bothered us, of course. I am not sure whether it would be better to have two separate machines for the latter purpose or one more complicated one. This depends on the degree of bi-uniqueness in the correlation. If there were a sufficient number of words in the FL to which only one word of TL corresponds so that to this word in TL only one word of FL corresponds, namely the same as before, then a combined machine might be worthwhile.
Wiesner: The tape on which the dictionaries are recorded do not form part of the machine. The machine proper, the computational part of it, could be the same for translation in either direction.
Bar-Hillel: No, this is not so. The grammatical analysis of different languages may be totally different. Therefore even for the analysis preceding the translation proper a two-way machine would, in general, be more complex than a one-way machine.
Bull: Why should we try to build a two-way machine?
Wiesner: I do not say that we should build such a machine. I only want to understand the underlying problems a bit better.
Reynolds: I think we may say roughly that a two-way machine would be l½ times as complex as a one-way machine. In addition to another dictionary, we would need additional inputs and outputs and a switching system to produce the reversal.
Bar-Hillel: Certain components of the machine could indeed be the same. But the part dealing with the operational syntax will just have to be twice as large.
Wiesner: Not necessarily. This would depend upon whether this part would be wired in or taped in.
Bar-Hillel: Indeed so. I do not know at the moment whether these operations will be performed by a set of instructions completely built in into the machine or by invoking instructions taped upon external devices.