This document is subject to the following notice:

NOTICE: This material may be protected by copyright law.

The copyright law of the United States (Title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material; the person using this equipment is liable for any infringement.
Why translation is difficult for computers

Doug Arnold
University of Essex, Colchester, England

Why is it difficult to get computers to translate? Our answer to this will be in two parts. The first part consists of some general remarks about the nature of translation, and the abilities of computers. These will lay out the ground and provide a general but rather unsatisfactory answer to the question. The second part will look in more detail at the sorts of problem that create the difficulty, and provide a more detailed and revealing answer.

1. Translation and computers

Part of the reason why translation is difficult for computers is that translation is just difficult: difficult even for humans. Translating is a many-faceted skill that goes well beyond mere competence in two languages. Roughly speaking, the job of a translator is to take a text in one language (the source language) and produce a text in another language (the target language) which is in some sense equivalent. Before we talk about why this is difficult, we should notice that translators are often asked to do rather more than this. In particular they are often expected to produce a text that is in some sense "good" in its own right — clear, unambiguous, interesting, persuasive, elegant, poetic, gripping, etc., according to the kind of text being translated. While this is understandable, it is clearly somewhat unfair, especially when one is thinking about trying to automate the process. It is one thing to ask a computer to produce a target text which is (in some sense) equivalent to the source text; it is quite another to ask the computer to make it interesting. So, in asking why translation is difficult for computers, we should be careful to restrict ourselves to the translation job proper: to be concrete, let us imagine that anything the computer produces will be post-edited for qualities other than equivalence with the source text. All we want from the computer is some kind of draft quality translation: something which is more or less faithful to the original, understandable in its own
right, and which is a reasonable starting point for a polished translation.

Of course, this is still very difficult, even for a skilled human, because the appropriate notion of “equivalence” is difficult to pin down, and can vary greatly depending on the kind of text involved. For example, in translating texts for an online help system, the length of the source text (number of characters) may be important, since the translation may have to fit in the same area of screen as the source text. While one normally expects a translation to be roughly the same length as the original, one would not normally worry about counting characters. Let us try to ignore these complications also, and focus on cases of translation where the key point is just to convey the content of the source text.

Unfortunately, this is still a tall order, because languages do not always allow the same content to be expressed. There are many well-known cases where one language lacks a precise equivalent for a term in another. In English, one can be vague about the gender of a friend, without seeming evasive. This is harder in French, where one has a choice between terms for male ami and female amie. Conversely, it is hard in English to refer to a friend who is female without going too far (girlfriend) or seeming to labour the point (female friend). So let us be a little less ambitious, and ask for only approximately the same content.

Even so, translating is a difficult task. In particular, it is a creative task, for at least two reasons. First, translators are often expected to be able to coin translations of novel terms that appear in the source text. Second, translators are often required to act as cultural mediators, conveying to readers of the target language what may be obvious to readers of the source language. A very clear case of this occurs with the translation of religious texts (how should one translate Man shall not live by bread alone for readers for whom bread is an alien or exotic foodstuff?)

Computers are fundamentally just devices for following rules, mechanically and literally, albeit with considerable speed and precision. Rule following can produce a kind of creativity, but not the kind of creativity required for these tasks. Coining a new piece of terminology is more a matter of inventing a rule than following a rule, and cultural mediation requires very sophisticated reasoning: one must be able not only to extract the meaning from a text, but also be able to think about what meaning a potential reader would extract. To avoid these problems, we should restrict ourselves to cases where readers of source and target text can be regarded as sharing the same culture and background knowledge (e.g. by being members of the same profession or scientific discipline), and where problems of novel terminology either do not arise or can be solved by a human in interaction with the computer.

The translation task we have now is one of taking a text written in one language and producing a text in another language with the same approximate content, where readers of the target text are expected to share the same knowledge and culture as the readers of the source text, where there are no problems due to new terminology, and where we expect a human translator to be involved in producing a polished result. For the most part, the aim of MT research over the last forty or so years has been to automate this process. Despite considerable progress, despite the fact that the aim has actually been achieved for some languages, and some restricted domains and text types, it still poses fundamental practical and theoretical problems.

At the root of these problems are four particular limitations of computers, namely, the inability of computers to:

(i) perform vaguely specified tasks
(ii) learn things (as opposed to being told them)
(iii) perform common-sense reasoning
(iv) deal with some problems where there is a large number of potential solutions.

Precisely formulated rules are required because they must, ultimately, be interpreted in terms of the normal operations of computer hardware. Much of the difficulty of natural language processing in general, and MT in particular, arises from the difficulty of finding sufficiently precise formulations of intuitively straightforward ideas like “in English, the subject usually comes before the verb” (the really problematic word here is usually, of course). Moreover, a precise formulation is not enough. There are problems for which rules can be formulated precisely, but for which solutions still cannot always be computed (any task that involves examining every member of an infinite set, for example).

Learning also poses fundamental problems from a computational perspective. There are several reasons for this, one of which is to do with the fact that it involves classification, which involves the notion of similarity, which is a vague notion, another being the fact that it involves genuine creativity (rule inventing, not rule following). There are learning algorithms for some tasks, but there is no general reliable procedure for learning the kinds of knowledge required for MT. In this area, what a computer needs to know, it must be told, in the form of explicit rules, written by humans.

The third problem is that computers cannot perform common-sense reasoning. There are several reasons for this, but perhaps the most serious is the fact that common-sense reasoning involves literally millions of facts about the world (water is wet, men don’t get pregnant, most people have two feet, sheep are larger than fountain pens, if B has been put in A then A contains B, for A to contain B, A must be larger than B, and so on). The task of coding up the vast amount of knowledge required is daunting. In practice, most of what we understand by “common-sense reasoning” is far beyond the reach of modern computers.

The fourth fundamental difficulty for computers arises even for precisely
Why translation is difficult for computers

Doug Arnold

specified problems which do not involve learning. It is the problem of combinatorial explosion. Suppose there are a number of slots each of which can be filled in one of two ways (say, by a 0 or a 1), and that we have to consider every way of filling the slots (the worst case). The number of possibilities very quickly becomes very big. There are two ways of filling one slot, four ways of filling two, and in general 2^n ways of filling n slots. Every time we add a slot, we double the number of possibilities, and hence the amount of time required. Suppose that it takes 1 millisecond to consider one solution: ten slots involves 2^{10} = 1024 possibilities, requiring just over a second. With 20 slots, the number of possibilities rises to 1,048,576, requiring over two hours. With 30 slots, the time goes up to 12 days, with 40 it goes up to over 34 years. Dealing with 41 slots would take over 64 years, which is too long for most humans to wait. Improvements to computer hardware are insignificant in the face of this sort of problem: buying a computer which is twice as fast as your present one allows you to deal with exactly one more slot in any given time.

The bad news, from an MT perspective, is that each of these limitations is relevant. Thus, a general, though not very revealing answer to the question we started with would be: “Because it involves problems that resist an algorithmic solution, including common-sense reasoning, learning, and combinatorially explosive tasks”. In order to give a more systematic and revealing answer, we need to look at the various tasks involved in different approaches to MT.

There are three “classical” architectures for MT. These, and the tasks they involve, can most easily be understood in relation to a picture like the well-known “pyramid diagram” in Figure 1, probably first used by Vauquois (1968).

The simplest approach to translation is the so-called direct approach. Here the aim is to go directly from the source-language text to a target-language text essentially without assigning any linguistic structure. Since no structure is assigned, translation has to proceed on a word by word basis. Examples where this goes wrong are all too easy to find, and we will have little more to say about the approach. A more promising approach is base on the so-called transfer architecture. Here translation involves three main tasks:

- Analysis, where the source text is analysed to produce an abstract representation or “interface structure” (IS) for the source-language text (IS_s). This typically contains some properties of the source language (e.g. the source-language words).
- Transfer, where the source-language representation is mapped to a similar representation of the target-language text (IS_t).
- Synthesis, or generation, where the target-language representation is mapped to a target text.

The third classical approach involves an interlingual architecture. Here the idea is that one has at one’s disposal an “interlingua”: a more or less language-independent representation scheme. The role of the analysis component is to produce an interlingual representation (IL), which the synthesis component can map to a target language text.

A simple way to understand the relationship between these approaches is to start with the three tasks involved in the transfer approach, and say that the interlingual approach tries to eliminate the transfer task, and the direct approach tries to do without analysis and synthesis (i.e. it reduces everything to the transfer task).

This division into three tasks provides a rough classification of problems for what follows. In outline, the more “revealing and systematic” answer which was promised will be in four parts:

- Form under-determines content. That is, it is not always easy to work out the intended content from what is written. This is the Analysis Problem (Section 2).
- Content under-determines form. That is, it is difficult to work out how a particular content should be expressed (because there is more than one way to say the same thing in any language). We will call this the Synthesis Problem (Section 4).
- Languages differ. That is, that there are irreducible differences in the way the same content is expressed in different languages. We will call this the Transfer Problem, since in a transfer-based system it is typically where this problem shows up (Section 3).
- Building a translation system involves a huge amount of knowledge, which must be gathered, described, and represented in a usable form. We will call this the Problem of Description (Section 5).
Basing discussion around the tasks involved in the transfer architecture in this way may invite the question of whether one could not avoid the problems simply by eliminating the corresponding tasks. We will say something about this in relation to interlingual approaches in Section 3 (where we will argue that though they reduce the “transfer problem”, they do not eliminate it), and in Section 5, where we will look at recent “analogical” approaches, and argue that though they offer partial solutions to these problems, the problems themselves remain.

2. The analysis problem

The task of an analysis component is to take a source-language text (e.g. a sentence), and produce an abstract representation — the idea being that it will be easier to translate from this representation than from an unstructured string of source-language words. There will be different views on what sort of representation this should be (e.g. how abstract it should be), but it clearly must represent the “content” of the source text, since this is what the source text and its translation have in common.

The problem is to infer the content from the source text. There are two major difficulties:

- The source text will often contain sentences that are ill-formed, at least from the viewpoint of the rules in an analysis component. Analysis components must be able to cope with this by being robust.
- The source text will often be ambiguous, so it may be difficult to work out what content is intended: the form of the input under-determines its content.

The problem of ambiguity is that no matter how superficial the representations we decide to use for an MT system, it will generally be the case that one string of words can correspond to several different representations.

The examples in (1) involve lexical ambiguity.

(1) a. They are trying to design a better pen. (‘writing implement’ or ‘animal enclosure’?)
   b. Our Jimmy has grown another foot. (‘limb’ or ‘unit of measurement’?)
   c. The post has arrived. (‘delivery of mail’ or ‘piece of wood’?)

The examples in (2) involve structural ambiguity — the indeterminacy of meaning is not due to any of the words, but to the different structures that can be assigned.

(2) a. Concern has been expressed about conditions in the factory near the river that was polluted last week.
   b. The minister stated that the proposal was rejected yesterday.
   c. Sam has joined a student film society.
   d. Some young girls and boys have arrived.

Is it the river, or the factory that was polluted in (2a)? What occurred yesterday in (2b), the rejection, or the minister’s statement? In (2c) is this a film society for students, or a society for student films (cf. adult film society)? Are the boys young, or is it just the girls in (2d)? The alternative interpretations of (2a) might be represented as (3).

(3) a. the [ factory near [ the river ] that was polluted last week ].
   b. the [ factory near [ the river that was polluted last week ]].

A very obvious and dramatic case of under-specification of content arises with pronouns, and other so-called anaphoric expressions. In an example like (4), one cannot tell who advocated violence: it might be the police, the women, or some other group that the speaker has mentioned earlier (or even a group that is being indicated in some other way).

(4) The police refused to let the women demonstrate because they advocated violence.

A legitimate question in relation to these examples is: “Does it matter?” There are no doubt many languages where a straightforward translation would preserve the ambiguity of some or all of the examples above. In these cases, surely the ambiguity should just be ignored.

One difficulty with this is that the cases that can be dealt with in this “ambiguity preserving” way are not the same for all languages. Example (4) about the police refusing the women a permit because they advocated violence is unproblematic for translation into languages with only one third-person plural pronoun, but it is a problem in relation to languages like French which make a distinction according to gender (ils vs. elles).

A second difficulty is that the cases where one needs to worry are somewhat unpredictable even for one pair of languages. One might, for example, think the structural ambiguity in an example like (5) could be ignored, when translating into French (is Pauline writing in Paris, or are the friends in Paris?).

(5) Pauline writes to her friends in Paris.

Indeed this ambiguity can be ignored with most verbs. But it must be resolved in translating a verb like miss (6), because its French translation, manquer, puts the noun phrase (NP) denoting the one who is missed in subject position, and realises the “misser” as a prepositional object. Thus, (6) has at least two non-equivalent translations, (7).

(6) Pauline misses her friends in Paris.
There are two key points to note regarding the automatic treatment of ambiguity. The first is that ambiguities combine to produce a combinatorial explosion. Consider a ten-word sentence, where each word is two ways ambiguous. Even before we consider structural ambiguities, or those with some other source, we have $2^{10}$ possibilities. Suppose there is a verb, followed by an NP, and a prepositional phrase (PP) (like example (6)). This gives an additional ambiguity, because the PP can be part of the NP, or not. So we may have as many as $2^{10} \times 2$ possibilities. If there is another PP the possibilities increase further. Now consider the pronoun her. It could, potentially be referring back to any female individual mentioned earlier. In the worst case, all these sources of ambiguity would be independent, and one is faced with a combinatorial explosion.

Fortunately, the worst case does not always arise because some of the ambiguities cancel out. In isolation either loves or presents can be a verb or a noun, but in (8), loves must be a verb, and hence presents must be a noun.

(8) Sam loves presents.

Nevertheless, in practice the number of possibilities is still very large, partly because most sentences are much more than ten words long, and most words are more than two ways ambiguous. It is reasonable to expect tens or even hundreds of analyses for quite ordinary sentences.

The second key point relates to the variety of information that would be required to disambiguate examples like these. Example (9) is very similar to (2a), but it is unambiguous because of grammatical information (the presence of a plural verb were unambiguously picks out the plural factories as the grammatical head of its subject, so the interpretation is factories ... were polluted). Thus, grammatical/structural information has a role to play.

(9) Concern has been expressed about conditions in the factories near the river that were polluted last week.

Similarly, (10) is unlikely to be interpreted as ambiguous, because of commonsense knowledge about the relative sizes of sheep and writing pens (and the fact that putting A inside B entails A being smaller than B):

(10) Sam put the sheep in the pen.

Likewise, young girls and boys is ambiguous, but pregnant women and men is not, because as a matter of fact, men do not become pregnant, and the reading one prefers for (4) will depend on all sorts of assumptions about women and police, and what constitute grounds for refusing permission to demonstrate.

In principle, it seems that any grammar fact or fact about the world, any piece of information or common-sense inference could be required for the correct resolution of an ambiguity.

Turning to the problem of ill-formed input, it is an unfortunate fact that ordinary written language, even the kind that has been carefully edited and prepared (like the contents of this book) abounds in errors of spelling, repeated words, transposed words, missing words, and what will appear to an analysis component to be errors of grammar.

Solutions (at least partial solutions) to these problems are not hard to find. For example, if we fail to produce an analysis for a whole phrase or sentence, we may nevertheless have successfully analysed parts of it, so we might try to hypothesize a missing word, or transpose a pair of words, and try to re-analyse, using the partial analyses that have been established. In a case like (11), we might just relax the requirement that a third-person singular subject requires a particular verb form. Of course, such tricks are a long way from the flexibility of the human reader, which is based on an overall understanding of the text.

(11) The problems are interesting, but the solution (sic) leave something to be desired.

However, two points should be kept in mind. First, inserting words, trying permutations of words and so on, are all potentially combinatorially explosive. Second, notice how dealing with ill-formed text interacts with the problem of ambiguity. The obvious way to deal with a case such as (11) is to disregard the rules that enforce subject-verb agreement. But doing this generally will lead to increased ambiguity. In particular, the unambiguous example (9) becomes ambiguous if one ignores the information about subject-verb agreement, because it becomes indistinguishable from (2a). In principle, this point holds for any restriction at all: imposing the restriction may lead to a failure in analysis; relaxing it will lead to more ambiguity.

All together, the problems posed by ambiguity and robustness may make the situation look rather desperate: reliable analysis seems to require nothing less than complete understanding at a level comparable to a human. Indeed, such considerations led some early researchers to declare that MT was not just difficult, but theoretically impossible. Fortunately, things are not quite this bad in practice. Partly, this is because, as noted, some ambiguities “cancel out”, and some can be
excluded by employing perfectly normal grammar rules (subjects and verbs agree in person and number). Restricting the domain and or text-type that one deals with will also be helpful. Some of the problems can be addressed by clever interaction with a human operator or post-editor (e.g. pronouns can be left flagged and left untranslated for a person to fix). If all else fails, one can just choose one interpretation, either at random, or on the basis of some ideas about which interpretations are more likely than others — this will be wrong some of the time, but most of the time it will be right.

3. The transfer problem

The task of a transfer component is to take the sort of abstract representation produced by the source-language analysis component (call this a "source IS"), and produce something that can be input to the synthesis component of the target language (call this a "target IS"). Obviously, the closer the two ISs, the easier this will be. The transfer problem is that they cannot be the same, because languages do not associate form and content in the same ways. Thus, rules must be written to relate source and target ISs.

To be concrete, let us assume that ISs are relatively superficial representations, along the lines shown in (12) and its translation (13).

(12) a. I miss London.
   b. [sentence/pres miss,
      [v/ing/1st PRO ],
      [np London ]]

(13) a. Londres me manque.
   b. [sentence/pres manquer,
      [v/ing/1st PRO ],
      [np Londres ]]

Given this sort of representation, the sort of thing transfer rules need to say is that London translates as Londres, that first-person singular NPs translate as first-person singular NPs (usually), and that translating a structure where the verb is miss involves getting the translation of its subject (the first NP) and its object (the second NP), and putting them in the appropriate slots in a structure whose verb is manquer, namely the indirect-object and subject slots respectively, and dealing with the tense, and so on.

The assumption is that though languages use different words and structures to express the same content, nevertheless there are enough similarities that words and structures can be put in some kind of fairly straightforward correspondence. Of course, this can easily lead to "translationese" where the structure of the source language is wrongly carried over to the target language. Nevertheless, the assumption holds and can be the basis of reasonable translation for many constructions and languages. Unfortunately, there are also many cases where the assumption fails.

Sometimes languages either package content differently, or just use radically different structures to express the same content. A case of the former can be seen in (14), which exemplifies a general difference between the way information about the direction and manner of motion is packaged in English and French (and other Romance languages). In English, the manner and motion are expressed in one item (the verb run), the direction is expressed by a PP (into the room). In French, the verb entrer 'enter' expresses motion and direction, while manner is expressed by an adverbial (en courant 'by running').

(14) a. He ran into the room.
   b. Il entra dans la chambre en courant.
   he entered into the room by running

Moreover, while it is possible to write a (rather complex) transfer rule that will state the correspondence here, this is in fact a quite general phenomenon, and it would be nice to have a general treatment, rather than dealing with individual cases (one rule for run into, one for walk into, one for fly out of, etc.)

A case of languages using radically different structures for roughly the same content can be seen in (15). Dutch (15a) involves a construction with an impersonal pronoun, Spanish (15b) uses a reflexive, and English (15c) uses a passive construction. If the corresponding IS representations are as superficial as those above, some very complex transfer rules will be required.

(15) a. Man verkoopt hier appels.
   one sells here apples
   b. Se venden manzanas aqui.
   self they-sell apples here
   Lit. 'Apples sell themselves here'
   c. Apples are sold here.

The need for very complex rules can also arise when two languages have corresponding constructions (i.e. content is packaged similarly), but the constructions are subject to different grammatical restrictions.

One example of this involves adjectives like difficult and easy and their translations in German. In (16a) the subject, Sam, is understood as one of the objects of the verb convince: compare (16b). The German (16c) is structurally parallel, and expresses the same content.
Unfortunately, there are differences between this construction in English and German. One difference is that while in English the understood position can be any kind of object, in German it must be a direct object. Thus, a straightforward translation of (17) produces the ungrammatical (18a). Instead, one must produce something like (18b), with a very different structure.

(17) Sam is easy to work with.
(18) a. *Sam ist einfach mitzuarbeiten.
   b. Es ist einfach mit Sam zu arbeiten.

It is important to notice that even apparently small differences between languages can give rise to problems. In English, the idea of being hungry is expressed with an adjective, in German a noun is used, as in (19).

(19) a. I am hungry.
    b. Ich habe Hunger.

Not much to worry about here, one might think: a rule to the effect that English hungry translates as German Hunger, should be sufficient. Sadly, this is not the case, as one can see from an example like (20) where the English adjective is “intensified” with a word like very. One cannot simply get the normal translation of very (sehr): instead the adjective intensifier very must be translated as a nominal intensifier viel ‘much’.

(20) a. I am very hungry.
    b. Ich habe viel Hunger.

Often these “knock-on” effects of the way content is expressed require information that is absent in one language to be supplied in another. A simple case of this arises in the translation of German examples like (21a) into English.

(21) a. das für Sam neue Auto
    b. the car which is new to Sam

The problem is that English does not allow nouns to be pre-modified in the same way as German (cf. *the new to him car). The solution is to make the modifying material into a post-modifier, putting it after the noun. This sounds easy enough, but moving the material after the noun involves turning it from a PP into a relative clause, and turning it into a relative clause involves supplying a verb (be), and when one supplies a verb one is also required to supply a tense (in (21b) we assumed it was present tense, but there is nothing in the German to indicate this).

A sort of limiting case of differences between constructions arises where one language completely lacks a construction, and one must always resort to finding a paraphrase. French simply lacks a resultative construction corresponding to (22a), so a straightforward translation is impossible. Instead of a simple adjective (flat), a whole subordinate clause is required, for which a tense, and a subject must be supplied (22c).

(22) a. They hammered the metal flat.
    b. *Ils ont martelé le métal plat.
    c. Ils ont martelé le métal jusqu'à ce qu'il est devenu plat.

Of course, the need to supply information that is unspecified in the source structure does not arise just because of particular constructions. It can arise between languages generally. For example, in English, one cannot avoid the issue of whether an NP is singular or plural, and whether it is definite or indefinite. In Japanese, on the other hand, this information can remain unspecified, so there is a clear problem in translating from Japanese into English. There is a similar problem going from English to Japanese, because in Japanese it is hard to avoid being precise about social relations between the writer and reader (e.g. it affects the form of the verb) which are not expressed in the English.

It is perhaps easy to see the general direction of a solution to these problems. The transfer problem arises because source and target language interface structures (ISs) differ. The more similar they are, the smaller the problem should be. Does this mean that one can avoid the problem entirely by adopting an interlingual approach? Unfortunately, it does not. The reason is that even under an interlingual approach it will be very difficult to ensure identity of source and target ISs.

First, however, it is worth noting a drawback to an interlingual approach, namely that making source and target representations more similar complicates analysis, by increasing ambiguity, sometimes unnecessarily. A simple example arises with languages (like Japanese and Chinese) which have different words for older and younger sister. This distinction will have to be present in an adequate interlingual representation for such languages. This means that producing an interlingual representation for English sister will involve disambiguation (older or younger sister?). This is entirely appropriate when the target language is Japanese or Chinese, but it is wasted effort when the target is another European language. (As
will become clear in the following section, adopting more abstract representations also complicates synthesis).

The following example will clarify the difficulty of ensuring identity of source and target representations. An approximate rendering of the content of English (23a) might be as in (23b), which says that for all events $e$, if $e$ is an eating event where the thing doing the eating is Sam, then the eaten object ($f$) is fish.

(23) a. Sam eats only fish.
   b. For all $e$: if [eating($e$) & eater($e$, Sam)]
      then [eaten-object($e$, $f$) & fish($f$)]

The same idea is expressed in Japanese as (24a), whose content is most naturally given as something like (24b), which says that “there are no eating events with Sam as eater that do not involve fish as object” (one reason for regarding this as a “natural” representation is that it correctly captures the negative nature of the Japanese sentence).

(24) a. Sam wa sakanashika tabenai.
   Sam topic fish apart eat-not
   ‘Sam does not eat anything apart from fish.’
   b. Not Exists $e$: [eating($e$) & eater($e$, Sam)]
   & not [eaten-object($f$) & fish($f$)]

Now these representations are equivalent. However, they are not identical, and it would clearly be difficult to find a general way of ensuring that this sort of thing does not arise. Not only would representations have to be very abstract, they would look utterly arbitrary from the point of view of some languages. (Why should Sam eats only fish involve a negative? Why should the Japanese not involve a negative?)

However, given the equivalence of these ISs, one might still hope to do away with transfer rules by formulating a general “inference” procedure along the following lines: take the source IS, input it directly to the synthesis component, if a correct target sentence is produced, then stop. Otherwise, find an equivalent IS, and try with that, etc. There are two worries here. First, it assumes we have a “logic” for ISs, which provides a well-defined notion of equivalence for ISs. Second, finding an equivalent IS is very likely to be one of the problems for which solutions cannot always be computed (because the number of equivalent ISs is likely to be infinite). It is, in any case, a combinatorially explosive process.

Thus, while using more abstract representations is clearly a good idea, because it will make transfer rules simpler, and while the transfer problem can be simplified by the right choice of representations, the implication of this argument is that there are irreducible differences in the way languages express “the same” content, and the transfer problem cannot be completely eliminated.

4. The synthesis problem

The two aspects of the synthesis problem are actually instances of the last problem discussed in the previous section. There are typically many ways in which the same content can be expressed. In short: meaning under-determines form.

The first aspect of the problem is that sometimes only one of the ways of expressing the content is correct. There seems to be no principled reason why (25a) is correct in English, rather than (25b,c).

(25) a. What time is it?
   b. How late is it?
   c. What is the hour?

On the face of it, these would be equally good ways of expressing the same content. It is just that only one is idiomatic English. The solution to this problem may look simple — just keep a list of the contents that must be realized by these semi-fixed expressions, and stop rules applying to produce the correct, but unidiomatic alternatives. But this solution is not foolproof, precisely for the reasons discussed at the end of the previous section: there are many ways in which the content that one would like to realize as (25a) could turn up in an IS representation, so it will be hard to list them all.

The second aspect of the synthesis problem is in some ways the converse of the first. It occurs when there is no obvious way of selecting the right way to express the content. To take a very simple example, the content of (26a) might be represented as (26b),

(26) a. Sam saw a black cat.
   b. Some $e$: seeing($e$), by($e$, Sam), of($e$, y), black(y), before($e$, now)
   i.e. there is a seeing event ($e$), where Sam did the seeing, and the seen thing ($y$) was a black cat, and the event occurred before now.

   This content can be expressed in English in many other ways (27).

(27) a. Sam saw a cat. It was black.
   b. Sam saw something black. It was a cat.
   c. Sam saw a cat which was black.
   d. Sam saw a black thing which was a cat.
   e. A black cat was seen by Sam.
   f. Something happened in the past. Sam saw a cat.
   g. There was a black cat. Sam saw it.
   etc.
3. Try to manage without explicit representations of linguistic (or non-linguistic) knowledge at all.

The first solution is attractive in theory, and has proved successful in practice (cf. the outstanding success of Météo — see Chapter 15), but its value is limited by the number of such domains that exist (it has proved very difficult to think of other domains that are as tractable as weather reports). The problem with the second solution is that existing dictionaries and grammars have normally been created with human users in mind, and so do not contain the kind or level of information required for use in MT. The third solution underlies one of the recent approaches which are discussed in the following section.

6. Other approaches

The preceding sections have looked at the problem of MT in terms of the “classical” approach, where translation takes place in three (or possibly two) stages, involving representations and explicit rules encoding various kinds of linguistic and other knowledge. The last decade has seen the emergence of so-called analogical approaches to MT, which, at least in their radical form, dispense with the representations and rules. The possibility arises that such approaches thereby solve some or all of the problems. This section will show why this is not the case, or at least why it is only partly the case. The analogical approaches in question are example-based approaches and stochastic or statistical approaches.

6.1 Example-based MT

The leading idea behind so-called Example-based MT (EBMT) approaches is that instead of being based on rules, translation should be based on a database of examples, that is, pairings of fragments of source- and target-language text (see also Chapter 3.6). Suppose, for example, that one has the pairings in (28) and (29) in the database, and has to translate (30) from English into French.

(28) I have a headache.
   J'ai mal de tête.
   I have ache of head

(29) I’d like something for a hangover.
   Je voudrais quelque chose contre la gueule de bois.
   I would-like some thing against the face of wood

(30) I have a hangover.
   J'ai la gueule de bois.

Ideally, what should happen is that matching (30) against the English parts of (28) and (29) will reveal that I have can translate as j’ai, and a hangover can translate as la gueule de bois, which can be combined together to produce the acceptable translation (31).

(31) J'ai la gueule de bois.

This is a very intuitive and appealing model of translation (for example, it seems to reflect something of the way humans work). In the best case, we may get an exact match for an input sentence, which will be paired with just one translation, and all the problems are solved. Unfortunately, in the general case things will not be so simple, and all the problems remain.

First, even when we have exactly matched an input sentence, it may correspond to several target examples, among which we must choose (if the database is sufficiently large and representative, genuinely ambiguous examples will get alternative translations). If the alternatives are equivalent, we have an instance of the synthesis problem of Section 4; if they are not equivalent, we have an ambiguity problem, analogous to the analysis problem of Section 2 (the difference is that we have to chose between alternative examples, rather than alternative representations). “Ambiguity” will arise in other ways, because there will typically be many other examples that partially match an input like (30), for example, those in (32) and (33). Each of these will suggest an alternative translation for I have, which will not yield correct translations (e.g. Je suis ‘I am’ and Je viens ‘I come’).

(32) I have left.
   Je suis parti.
   I am departed

(33) I have just left.
   Je viens de partir.
   I come of to-depart

Moreover, given that (28) has been chosen as a match for part of the input (30), we have to decide which parts of the French translation to take: how do we decide that j’ai corresponds to I have? This is like the transfer problem of Section 3 in that it will be harder to work out correspondences the more source and target examples diverge from word-for-word alignment (i.e. the more the languages diverge in the way they express content). Finally, having decided which pieces of the French examples we need to combine, how do we decide to combine them as in (31), rather than in the other order, or somehow mixed up? This is again somewhat analogous to the synthesis problem of Section 4.

In principle, one might still hope to manage these problems without recourse to rules (i.e. explicit linguistic knowledge). For example, one might observe that the
sequence *ai mal 'have ill(ness)' occurs much more frequently than *viens de mal 'comes of ill(ness)', and on this basis choose the correct (31) over the incorrect (and meaningless) (34).

(34) *Je viens de la gueule de bois.

However, this leads us directly to the EBMT version of the problem of description, which is that in order to make the approach work one will need many millions of examples. Bilingual dictionaries provide one source of appropriate examples, but this will yield at most a few thousand. For the rest, we must rely on "aligned corpora", of which we will say more shortly.

6.2 Statistical approaches

The intuitive appeal of statistical approaches can be seen when one considers how one normally approaches very complex processes involving a large number of interacting factors. One approach is to try to disentangle the various factors, describe them individually, and model their interaction. One might, for example, try to model the way a crowd behaves by trying to understand how every individual in it behaves, and how they interact. But for many purposes, including the case of crowd behaviour, it is more sensible to step back and try to model all or part of the process statistically. Given that translation is a very complex process involving many factors, the appeal of some kind of statistical methodology should be clear.

Of course, there are many ways one could try to apply statistical methods in a "classical" approach to MT, but a more radical idea has also been proposed. The central idea is this. When presented with a French sentence $f$, we imagine that the original writer actually had in mind an English sentence $e$, but that $e$ was somehow garbled in translation so that it came out as $f$. The job of the MT system is just to produce $e$ when presented with $f$. Seen in this way, translation is an instance of transmission down a noisy channel (like a telephone line), and there is a standard technique that can be used to recover the original input (the English sentence $e$), at least most of the time. The idea is that $f$ is more or less likely to occur depending on which English sentence the writer had in mind. Clearly, we want the one(s) that give $f$ the highest probability. Moreover, it also makes sense to take into account the relative probabilities among the English sentences (perhaps the probability of getting (35a) given (35b) is not much different from that given (35c) but the former has a higher probability, and is the right choice of course).

(35) a. Quelle heure est-il?
b. What time is it?
c. What hour is it?

To make this work, we need: (a) a statistical translation model, which assigns a probability to $f$ for each sentence of English; (b) a monolingual statistical model of English, which assigns a probability to every sentence of English; and (c) a method for finding the best candidate for $e$ according to these models.

Notice that since in principle any English sentence could give rise to $f$, (c) is a combinatorially explosive problem par excellence (even if we restrict ourselves to sentences that are about the same length as $f$, there will be millions of possibilities), but if we can find a (presumably imperfect) way of searching through the vast number of possibilities, we have a method that works without rules (hence no problem of description), without analysis (no problem of robustness — even completely ungrammatical sentences have some probability, however small), without intermediate representations (hence no problem of ambiguity deciding which representation to assign), and no problem of synthesis (deciding which sentence to produce given a particular representation).

Sadly, there are two important reasons why this is not a panacea. The first relates to a different version of the problem of description. The second relates to the quality of the available statistical models.

The statistical version of the problem of description is the problem of sparse data. Consider just the model of English: the only way that we can be sure that, say, (35b) is more probable than (35c) is by analysing huge amounts of text, and seeing that time appears more often in this context than hour. The problem is that in order to do this for most expressions we will need to examine astronomically large amounts of text. Even if one looks at many millions of words, many words appear only once or twice. So, there is a real problem getting reliable statistics. The problem is worse still when one considers translation. Here one relies on aligned parallel corpora, that is, collections of texts in two languages which are supposed to be translations of each other, which have been aligned so that it is simple to find the target-language sentence(s) that translate any particular source-language sentence. The classic example is the Canadian Hansard Corpus, consisting of reports of proceedings in the Canadian Parliament, which are published in both English and French. But such corpora are rare (non-existent for many pairs of languages), and tend to be relatively small. And of course the translation model typically needs more data than the monolingual model (whatever the probability of seeing an expression on its own, the probability of seeing it as the translation of some other expression must generally be lower).

A standard example of a monolingual statistical model is a so-called bigram model like those which have been very successfully applied in speech recognition. They involve the simplifying assumption that the probability of any given word sequence can be identified with the joint probability of each word occurring, given that the preceding word occurred. The probability associated with *The cat died is
the joint probability of *The* occurring as the first word in a sentence, of *cat* occurring given that the preceding word was *The*, and the probability of *died* occurring given that the preceding word was *cat*. The basic data for such a model is thus observations about the relative frequency of various pairs of words (bigrams). A generalization of this for the translation case might assume that the probability of *f* appearing as the translation of word *e* depends on the predecessors of *f* and *e*. But of course, it is clear what is wrong with this model. While the probability of *cat* clearly is influenced by the probability of *The*, this is not because *The* is the word before *cat*, but because *cat* is a noun, and nouns appear in NPs, which often start with determiners (like *The*). For example, *The* is exerting the same sort of effect in an expression like *The big cat* or *The big fat black cat*, while in a bigram model the effect falls off dramatically (it is seen as depending on the likelihood of *big* following *The*, and *fat* following *big*, and *cat* following *fat*). Of course, we can replace the bigram model with one that takes account of the grammatical structure, but now we are back with at least some of the ambiguity problems again, because taking account of the grammatical structure will involve linguistic rules and representations. Moreover, the statistical version of the description problem will be worse, because we need statistics not just about what words come next to each other, but about what structures go with what strings of words. Such statistics will be hard to find, because they will require text that has been analysed and given a representation (and giving text the right representation takes us straight back to the problem of ambiguity again).

Notice that the value of statistical methods *per se* is not at issue here, because to use statistical methods it is not necessary to adopt such a radical stance. One might, for example, try to use such methods to achieve robustness and disambiguation in analysis (e.g. if one encounters a lot of finance-related words in a text, it is quite likely that an occurrence of *bank* will denote a financial institution). But they are not a panacea, because a statistical method is only as good as the statistics, and these depend on what factors the model takes into account, and on the amount and quality of the data. Once one accepts the need for abstract representations, one is immediately and inevitably faced with the problems discussed in the rest of this chapter. Statistical methods are a contribution to the solution, but they are not in themselves the solution one might have hoped.

7. Conclusion

In short: translation is about producing a target text that preserves the content of the source text. This is difficult for computers because (a) form under-determines content; (b) content under-determines form; (c) languages differ in the way they express content; (d) it is difficult either to express the principles involved with the necessary precision, or to find the data needed for a statistical approximation.

Further reading

The reader who wants more background, details, or other perspectives can find extended, but still introductory, discussion of the issues discussed here, as well as references to the primary literature in, *inter alia* Arnold et al. (1993), Hutchins and Somers (1992), Kay et al. (1994) and Whitelock and Kilby (1995), and elsewhere in this volume, of course. More advanced and technical discussion can be found in Trujillo (1999).

Notes

1. This is much simpler than the task we started out with, which is one reason that MT is not a serious threat to the employment prospects of human translators.
2. In fact, this is not a new problem: it is just an instance of the ambiguity problem discussed above. The difference is that questions such as whether writer and reader are being polite or familiar only appear to be ambiguities when one thinks about translation into a certain language. But as with resolving ambiguity in analysis, inferring the necessary information can require nothing less than full understanding of the text (and context).
3. For example, one may want analysis to accept ungrammatical inputs, which one does not want synthesis to produce.

References


1. Introduction

In this chapter we consider the various ways in which linguistics, the scientific study of language, can be exploited in MT systems. Linguistics can be used by MT system developers in a rather random way, as a source of analyses or ideas for solving a specific problem. The earliest, “direct” systems may be said to have at best taken this kind of approach, and at worst to have ignored linguistic findings altogether. In this chapter, however, we shall be concerned with more rigorous and systematic uses of linguistics.

Linguistics is concerned with providing descriptions of languages, theories of human language in general, and formalisms within which these descriptions and theories can be stated. Linguistics is relevant to human translation as well as MT, though we will be covering mainly the latter here.

This chapter is structured as follows. In Section 2 we present an overview of approaches to linguistics and of how MT can make use of this discipline. In Section 3 we examine ways in which linguistics can contribute to defining the abstract representations used in MT, and in Section 4 take this further by looking at examples where source and target sentences differ considerably in structure. Section 5 studies the translation of tense, and Section 6 deals with functional aspects of language and the extent to which MT systems can incorporate these. Section 7 is a brief conclusion.

2. The field of linguistics

This section examines some of the ways in which linguistics can feed into MT research. To assist the reader, we can summarise the main contrasts we shall set up as follows: