CHAPTER 28

# MACHINE TRANSLATION: LATEST DEVELOPMENTS

HAROLD SOMERS

**Abstract**

The chapter reviews the current state of research, development, and use of machine translation (MT) systems. The empirical paradigms of example-based MT and statistical MT are described and contrasted with the traditional rule-based approach. Hybrid systems involving several approaches are discussed. Two recent developments within the rule-based paradigm are discussed, namely, anaphora resolution for MT and interlingua- and knowledge-based MT. As a major new application, spoken language MT is introduced. The prospect of MT systems for minority and less-developed languages is discussed, along with the use of MT on the Internet, notably for web-page translation. Finally, tools for translators are described, particularly those which exploit bilingual parallel corpora (translation memories, bilingual concordances), as well as translator-oriented word-processing tools.

## 28.1 Introduction

This chapter follows on naturally from the previous one, concentrating on the most recent research themes, and the current state of play in actual use of machine translation (MT) systems in the real world. The past ten years have been a very fruitful period for MT research and development on all fronts.

For basic research, the 1990s were marked by the emergence of a fairly new paradigm to challenge and eventually enrich the established methodologies. This was the so-called *empirical* approach, based on increasingly available amounts of 'raw data' in the form of **parallel corpora**, i.e. collections of texts and their translations. As we will describe below, these newly available data suggested to some researchers a way to approach the problems of MT which differed significantly from the traditional linguistic rule-based approaches. In parallel, several new ideas emerged under various names, notably **example-based MT** but also *analogy-based*, *memory-based*, and *case-based* MT, all having in common the use of a corpus or database of already translated examples, and involving a process of matching a new input against this database to extract suitable examples which are then recombined in an analogical manner to determine the correct translation.

A slightly different, but still corpus-based, approach has been called statistical MTit is clearly example based in that it depends on a bilingual corpus, but the translation procedure depends on statistical modelling of the word order of the target language and of source–target word equivalences. There is thus a focus on the mathematical aspects of estimation of statistical parameters for the language models.

Not all the latest research in MT has been in this 'new' paradigm. Much work is continuing within the more traditional **rule-based** paradigm, often on new language pairs, sometimes attempting to address hitherto more difficult problems. Typical of this is work on **anaphora resolution** in the context of MT, and on the development of ontological knowledge in the context of **interlingua**-based MT. New directions in computational linguistics in general have found applications in MT, such as the development of *tree-adjoining grammars* (TAGs), which specify relationships between different representations, and so have a very clear application in transfer-based MT. Another is the lexicalist approach known as *Shake and Bake* MT which shuns explicit analysis and transfer rules in favour of a constraint- and unification-based approach.

An important new development for MT in the last decade has been the rapid progress that has been made towards developing **spoken language MT** systems. Once thought simply too difficult, improved speech-analysis technology has been coupled with innovative design to produce a number of working systems, albeit still experimental, which suggest that this may be the new growth area for MT research.

Finally, we will mention arguably the most significant development to have influenced MT research, namely the *Internet* and the use of MT by web surfers. This ori-

ginally seemed to take MT researchers a little by surprise, but the success of a number of MT systems which are available on the World Wide Web, usually free, has introduced an essentially new and unforeseen use for **low-quality MT**, as well as heightening awareness of MT for the general public. It has also introduced a number of new translation problems however, which we will discuss below.

## 28.2 Empirical Paradigms

### 28.2.1 Example-based MT

**Example-based MT** (EBMT) was first proposed as long ago as 1981 (Nagao 1984), but was only developed from about 1990 onwards. The basic idea is to reuse examples of already existing translations as the basis for a new translation. In this respect it is similar to (and sometimes confused with) the translator's aid known as a **translation memory** (TM) (see below). Both EBMT and TM involve matching the input against a database of real examples, and identifying the closest matches. They differ in that in TM it is then up to the translator to decide what to do with the proposed matches, whereas in EBMT the automatic process continues by identifying corresponding translation fragments, and then recombining these to give the target text.

The process is thus broken down into three stages: matching (which EBMT and TM have in common), alignment, and recombination.

The *matching* stage can be implemented in a variety of ways, depending crucially on how the examples are stored in the first place. In early EBMT systems, examples were stored as fully annotated tree structures, with explicit links between the constituents in the two languages. The new input would be parsed, using the same grammar as the one used to build up the example database, and the resulting tree compared with the trees in the example database. Lexical differences are quantified using a hierarchical thesaurus. Since these trees were already aligned, all that remained was to cut and paste the partially overlapping tree structures in a fairly simple way. This arrangement works quite well, but the computational overhead is quite significant. In particular, the need for a traditional rule-based grammar with which to parse the input, and the need to align the tree structures (usually manually), meant that this approach to EBMT only really differs from traditional MT in the way it replaces the transfer stage, and so it is often referred to as *example-based transfer* rather than EBMT proper.

A more radical approach is to treat the examples, and the new input, just as strings of characters. The matching part of the process is then an example of *sequence comparison* found in many other computational applications, and for which several suitable algorithms exist. Again, a thesaurus can be used to quantify lexical substitutions,

though other measures have also been used. Because there is no tree structure to rely on, the alignment and recombination phases become much more complex in this scenario.

In between these two extremes are approaches in which the examples are annotated to a greater or lesser degree. Quite widespread is the use of *POS tags* (see Chapter 11). Another approach is to combine several similar examples into a more general single example containing variables. Some systems use all of these approaches, so that the example set is a mixture of some very general cases which are effectively like the rules in rule-based MT, some partially generalized cases, and some literal examples.

One factor which applies in all of these cases is the *suitability* of examples. In the most purist of approaches to EBMT, the examples will be drawn from a real corpus of already existing translations. But such collections contain 'noise' in the form of *overlapping* or *contradictory* examples. Some researchers address this problem by eliminating troublesome examples, or changing them, or, in the extreme case, hand-picking or creating the examples manually. This approach has been criticized as repeating the major difficulty with rule-based MT, namely the reliance on hand-coded linguistic rules which are subject to human frailties such as inconsistency and whim.

The matching stage finds examples that are going to contribute to the translation on the basis of their similarity with the input. The next step is to identify which parts of the corresponding translation are to be reused, referred to as *alignment*. This will be straightforward if the examples are stored in a way that makes the links between the texts explicit, but otherwise it may involve some processing, perhaps using a bilingual dictionary, or comparison with other examples. This alignment stage is carried out by humans in the case of translation memory systems (see below), but must be automated for EBMT (see example (28.10) below). In some systems, the matching stage will identify several suitable examples each containing parts of the text to be translated. For instance, if we wanted to translate (28.1*a*) on the basis of retrieved examples (28.1*b*, *c*) (examples are from Sato and Nagao 1990), we would have to be able to identify which portions of the Japanese equivalents correspond to the underlined text in examples (28.1*b*, *c*).

(28.1) *a*. He buys a book on international politics.
*b*. He buys a notebook. *Kare wa nōto o kau.*
*c*. I read a book on international politics. *Watashi wa kokusai seiji nitsuite kakareta hon o yomu.*

This brings us to the third stage, called *recombination*. Having identified which parts of the examples to reuse, we have to be sure that we are putting them together in a legitimate way. We can illustrate this by considering German, a language which has explicit case marking to distinguish subjects and objects. Suppose we want to translate (28.2*a*) on the basis of the examples (28.2*b*, *c*). The German text corresponding to the phrase *the handsome boy* is different in each example. We would have to know something about German grammar to know which alternative to choose.

(28.2) *a*. The handsome boy entered the room.
*b*. The handsome boy ate his breakfast. *Der schöne Junge aß seinen Frühstück.*
*c*. I saw the handsome boy. *Ich sah den schönen Jungen*.

### 28.2.2 Statistical MT

In its pure form, the statistics-based approach to MT makes use of no traditional linguistic data. The essence of the method is first to align phrases, word groups, and individual words of the parallel texts, and then to calculate the *probabilities* that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language. An essential feature is the availability of a suitable large bilingual corpus of reliable (authoritative) translations.

This approach is often seen as 'anti-linguistics' and is most closely associated with the IBM research group at Yorktown Heights, NY (Brown et al. 1990), who had had some success with non-linguistic approaches to speech recognition, and turned their attention to MT in the early 1990s. As already mentioned, the idea is to *model* the translation process in terms of statistical probabilities: to use their example, if we take the input sentence (28.3), then amongst the *possible* translations are (28.4*a*) and (28.4*b*).[1]

(28.3) President Lincoln was a good lawyer.

(28.4) *a*. *Le président Lincoln était un bon avocat*. ['President Lincoln was a good lawyer']
*b*. *Le matin je me brosse les dents*. ['In the morning I brush my teeth']

What should emerge is that the probability that (28.4*a*) is a good translation is very high, while the probability for (28.4*b*) is very low. So for every sentence pair *S* and *T* there is a probability P(*T*|*S*), i.e. the probability that *T* is the target sentence, given that *S* is the source. The translation procedure is a question of finding the best value for P (*T*|*S*).[2]

This probability calculation depends on two other measures. The first is the probability that the words in *S* will 'give rise to' the words in *T*, the so-called *translation model*. The second is the probability that these words are correctly combined in the target language, which we could call the (*target*) *language model*.

Probability that a given word in the source text is 'responsible' for a given word in the target text, $P(w_t|w_s)$, can be calculated on the basis of an **aligned parallel corpus**. This is a laborious computation, but one which can be done once and for all for a given parallel corpus, looking at the relative distribution of all the words in the corpora. So

[1] English translations shown in square brackets are for the benefit of readers, and are not part of the MT system.

[2] The original article presents this the other way round, modelling $P(S|T)$, and calculating the probability of the source sentence given the target. We prefer to describe the approach in a more intuitive manner, which is nevertheless faithful to the underlying approach.

for example, it might appear from the corpus that the probability of the English word *the* being translated as *le* is 0.610, as *la* 0.178, and so on. An added complication in this calculation is the fact that word correspondences are not 1:1, the problem of *fertility*. For example, the English word *not* corresponds to *ne* and *pas* with probabilities of 0.469 and 0.460 respectively. The probability that *not* has a fertility value of 2 is quite high (0.758), so together this gives a good probability that *not* corresponds to *ne . . . pas*. For an interesting illustration of how this works consider the translation of the English word *hear*. The IBM group experimented with the English–French translations of the Hansard corpus of Canadian parliamentary proceedings. In this corpus, *hear* is translated as *bravo* with a probability of 0.992, but with fertility probabilities almost equally divided between 0 and 1. In other words, *hear* is only translated about half the time, but when it is, it comes out as *bravo*. This strange state of affairs is understandable when one considers how often in the corpus the English phrase *hear hear* is translated simply as *bravo*.

The translation model, then, gives us a 'bag' of the words most likely to be in the translation. The second part of the probability calculation must determine what is the best (i.e. most probable) arrangement of these words. This is the (*target*)-*language model*, which consists of probabilities of sequences of words, based on the model corpus. Bearing in mind that each source word is responsible for a number of possible target words, each with an associated probability, it becomes apparent that calculating the probabilities of all the possible sequences of all the possible target words is a huge task. Fortunately it can be simplified somewhat, since the probabilistic effect of sequence becomes infinitesimally small as the distance between the two words increases. It is sufficient therefore to calculate the probabilities of relatively short sequences of words, called *n*-grams, where *n* is the number of words in the sequence. Bigrams, for example are sets of word pairs, trigrams consist of three-word sequences. Yet even with this simplification, a large calculation is involved. So a useful starting point is to assume that the target-word sequence is the same as the source-word sequence. But we know that this is not generally true of language translation: there is often a degree of *distortion* of the target language word order with respect to the source language. By allowing a certain amount of distortion, the search space can be limited, and probabilities of the most likely target sentences considered.

What is striking about this approach is the complete lack of linguistic 'knowledge' used in the process. If the system 'knows' that *la* and *table* go together, it is because it has seen this combination most often, not because it knows anything about gender agreement. When first reported, researchers using rule-based methods were surprised that the results were so acceptable: almost half the phrases translated either matched exactly the translations in the corpus, or expressed the same sense in slightly different words, or offered other equally legitimate translations. Nevertheless, as research progressed it became apparent that the possibilities of improving the performance of these systems were very limited. In particular, many of the errors made by the system could be corrected with the most minimal (and non-contentious) linguistic knowledge, such as notions of morphology, agreement, and so on. Subsequently, the IBM

group made attempts to incorporate such linguistic knowledge—although remaining true to their empirical approach in that such information was always based on statistical observation—but the results were not extremely promising, and the group disbanded in the late 1990s.

### 28.2.3 Hybrid Approaches

Neither the example-based nor the statistics-based approaches to MT have turned out to be demonstrably better than the rule-based approaches, though each has shown some promise in certain cases. As a result of this, a number of **hybrid approaches** quickly emerged. Recognizing that some specific problems were particularly suited to an example-based approach, in some systems there is an example-based component which is activated specifically to deal with the kinds of problems that are difficult to capture in a rule-based approach. Other hybrid systems combine rule-based analysis and generation with example-based transfer. A third combination seems particularly suited to the thorny problem of spoken language translation, where for example elements of the analysis part may rely more heavily on statistical analysis, while transfer and generation is more suited to a rule-based approach.

A rather different type of hybrid is the case of **multi-engine systems**. In this case, the source text is passed through a number of different MT systems, each using different techniques. One may be essentially lexicon based, another rule-based analysis and generation, a third example based or more purely statistical. In each case, built into the system will be a kind of scoring mechanism, by which the engine is able to evaluate for itself its 'confidence' in the output. For example, a rule-based engine may be able to reflect how sure it is of having been able to choose correctly between competing analyses. At the other end of the process is a kind of 'moderator' which will take the outputs of the various engines and compare them, choosing the highest scoring proposal, or confirming similar translations proposed by different engines, or perhaps even consolidating them by combining the best bits of each.

## 28.3 Rule-Based Approaches to MT

### 28.3.1 Anaphora resolution

The interpretation of **anaphora** (i.e. coreference especially, for example, pronouns, see Chapter 14) is crucial for the successful operation of MT. This is particularly

evident when translating into languages which mark the gender of pronouns, or from languages which have *zero-anaphora* constructions into languages where the pronouns must be inserted. A further problem is that anaphoric reference often crosses a sentence boundary, whereas most MT systems are limited to the sentence as a translation unit. The problem is at its most acute when the system is used to translate conversational texts which are especially rich in anaphoric devices. In addition to anaphora resolution itself (i.e. the identification of links between anaphoric expressions and their antecedents) being a very complicated task, translation adds a further dimension to the problem in that the reference to a discourse entity encoded by a source language anaphor by the speaker (or writer) has not only to be *identified* by the hearer (translator or translation system) but also *re-encoded* as a coreferential expression in a different language. For example, *elle* is used in French to refer to a grammatically feminine noun, for which the appropriate translation may be *it* rather than *she* as in (28.5).

(28.5) *L'eau est claire mais elle est froide.*
'The water is clear but it (*she) is cold.'

In recent years there has been a growth of research in this area, covering a number of languages including Japanese, German, French, Portuguese, Chinese, Spanish, Bulgarian, Italian, Russian, Polish, Arabic, Swedish, Turkish, Hindi, and Malayalam. Obvious techniques such as recognizing number (and gender) concord (e.g. in English *they* will usually be linked to a plural antecedent, *she* to a singular female one) can be supplemented by *heuristics* reflecting the most likely clues. For example, parallel structures often suggest a link as in (28.6), where at least *video recorder* and *red button* are possible antecedents for the pronoun *it* (and there may be others, from earlier sentences).

(28.6) To turn on the video recorder, press the red button. To program *it*, press the 'Program' key.

Apart from heuristics like these, one way to resolve anaphora ambiguities is by better 'understanding' of the text. For example, in (28.7*a*), knowing whether video tapes or recorders are more likely to be rewound tells us which is the correct link for the pronoun, whereas in (28.7*b*) *it* refers to the machine.

(28.7) *a.* Insert the video tape into the recorder, rewinding *it* if necessary.
*b.* Insert the video tape into the recorder, after making sure that *it* is turned on.

Sometimes, pronouns refer to items only *implicit* in the text, in which case we need to understand the underlying situation, as in (28.8), where *it* refers to a not-mentioned meal or food.

(28.8) We went to a restaurant last night. *It* was delicious.

These approaches imply the need for richer linguistic information in the system,

which may be provided by incorporating *ontological knowledge bases* into systems. This is the theme of a second strand of research within the rule-based paradigm.

### 28.3.2 Interlingua-based MT

Dismissed in the 1980s as largely impractical, the notion of **interlingua**-based MT has undergone a revival in recent years. An example of this approach is the **knowledge-based MT** (KBMT) research based at Carnegie-Mellon University and NMSU. The *text-meaning representation* (TMR) in their multilingual MT system represents the result of analysis of a given input text in any one of the languages supported by the system, and serves as input to the generation process. The meaning of the input text—derived by analysis of its lexical, syntactic, semantic, and pragmatic information—is represented in the TMR as elements which must be interpreted in terms of an independently motivated model of the world (or *ontology*). The link between the ontology and the TMR is provided by the lexicon, where the meanings of most open-class lexical items are defined in terms of their mappings into ontological concepts and their resulting contributions to TMR structure. Information about the non-propositional components of text meaning—pragmatic and discourse-related phenomena such as speech acts, speaker attitudes and intentions, relations among text units, deictic references, etc.—is also derived from the lexicon, and becomes part of the TMR. The approach is made tractable by restricting the system to specific domains and by adopting the *controlled language* (see Chapter 22) approach to syntactic coverage.

The ontology at the heart of the system is a model that a particular speaker has about the world; this model is populated by *concepts*, organized in a particular hierarchy. The concepts in the ontology cover things (such as aeroplanes, ideas, or giraffes), events (e.g. buying, eating), as well as relations. The ontology is organized as a *taxonomy*, e.g. a concept such as HAMMER would be identified as a type of HAND-TOOL, with more specific types of hammer connected to HAMMER. As well as this basic IS-A-TYPE-OF link, other relations can be encoded in the ontology, such as HAS-AS-PART, IS-AN-OCCUPANT-OF, MANUFACTURED-BY, and so on, depending on the domain. The ontology can also be extended by associating frame-like information with concepts, such as COLOUR, SIZE, OWNER. Events in the ontology have associated case slots like AGENT and LOCATION, which in turn might have information about associated typical fillers.

The ontology, as mentioned above, is associated with a (multilingual) *lexicon*. Analysis and generation components are also of course necessary, but, as is usual in an interlingual system, no transfer component. Thus, the analyser can be used to generate a TMR for a text, and from this the target language text can be generated

directly. One implementation of this architecture translates Spanish news articles into English, but other languages are also covered.

## 28.4 Spoken Language MT

As recently as fifteen years ago, the task of MT applied to spoken language was thought too difficult for all but the most basic research. Recent developments in speech processing, coupled with new ideas about MT (EBMT and statistical approaches, as described above) have meant that **spoken language translation** (SLT) is now a major research avenue within MT.

It might be thought that SLT was simply a matter of coupling a speech-to-text front-end and a text-to-speech back-end to a conventional text MT system, but this approach would be completely inadequate for all but the most formal types of spoken language. Spoken language is hugely different from written language, apart from the obvious difference of medium (sound vs. text), which involves an amount of pre-linguistic processing to isolate the speech signal from the surrounding background 'noise'. Among the problems particular to SLT, depending on the type of speech, are identifying and processing spoken language phenomena such as hesitations and self-repairs (some of which actually serve a subconscious pragmatic function); correctly interpreting speech act phenomena and discourse functions; dealing with different accents, and mixed-language speech; much greater use of anaphora and ellipsis; ill-formed utterances, or rather, varied grammaticality of spoken language.

Interestingly, the field is dominated by one sort of SLT system, aimed at translating dialogues, and more particularly *cooperative dialogues*, for example between a traveller and a travel agent, where the dialogue partners collaborate towards a common goal, as opposed to *adversarial* dialogues, e.g. between business persons negotiating a contract. The implications of this apparently minor distinction are quite enormous, especially in terms of interpreting the pragmatic aspects of the dialogue. Other distinctions that might impinge on the design of an SLT system include

- whether it is face to face or telephonic;
- whether it has the possibility of *interactive disambiguation* and/or confirmation, and if so ...
- whether this also is speech based (introducing the difficulty of identifying system–user *metadialogue*) or on a separate user interface;
- whether users are purely monolingual or may switch languages from time to time.

Almost all the SLT literature focuses on *dialogue translation*, with very little work as yet reported on what might be termed, by analogy with MT, *machine interpretation*, that is, simultaneous or consecutive translation of spoken language in the context of a meeting or a person addressing a group of people. Interestingly, this might prove to be a somewhat less difficult task than dialogue translation, apart from the exigencies of real-time processing of course, since the type of language that gets interpreted (by human interpreters) is usually much more formal than everyday dialogue, and closer in nature to the written language. Another application that does not yet seem to have attracted much attention is the SLT corollary of email translation, namely *voice-mail*, where there is also scope for translation. The problems would be similar to those involved in translation of other spoken messages, for example between emergency or security services across linguistic borders, e.g. in the Channel Tunnel.

## 28.5 Development and Use

### 28.5.1 MT for minority languages

A recent area of activity in MT and related fields cuts across the research, development, and use boundaries. This is the application of language technologies to less-favoured languages, generally (though perhaps misleadingly) termed *minority languages*. This term is misleading, because it refers not only to languages with small numbers of speakers, like Welsh, Breton, and so on, but also to languages which, because of the economic and geographic situation of their speakers, have not received much attention. Among these are languages with the most numerous speakers in the world: Urdu, Hindi, Cantonese, for example. These are of interest also in the 'West' as **non-indigenous minority languages** (NIMLs). These languages have not caught the attention of researchers and developers until now, for obvious, mainly economic, reasons: attention has been focused on European languages, plus Chinese (i.e. Mandarin), Japanese, Korean, Arabic. Now there is a small but growing area of activity to promote the development of MT, or at least related tools, for people using these minority languages. Fortunately, advances have been made at least regarding basic tools such as fonts and character sets, without which of course almost nothing of much use could be achieved. But beyond that there is a huge amount of work involved in building up grammars and lexicons, structured vocabularies, terminology, and so on for so many languages.

One approach that has been advocated is to look at research in **corpus linguistics**, where experiments have been reported in which linguistic information (word lists, rudimentary grammars,and even bilingual lexicons) can be extracted semi-auto-

matically from mono- or bilingual text corpora. It remains to be seen to what extent these techniques, largely developed for European languages, can be extended to typologically varied languages; but at least the raw material, in the form of text corpora, is becoming more and more widely available, as the World Wide Web grows. It has been suggested that English is no longer a 'majority' language on the web, and the only impediment to the growth of linguistic diversity on the web seems to be the availability of computers, and, to a certain extent, standardization.

Related to this is research on rapid development of MT for new language pairs, perhaps within an existing computational framework. This has been supported largely by political motivations, whereby the language needs of the major powers fluctuate depending on sociological (and sometimes naturally occurring) events around the world. Researchers have been looking into methods which will enable language technology tools ranging from on-line dictionaries and phrase-books through to computer-aided translation and full MT systems to be developed quickly, to support aid workers, military and political advisers, and various other interested parties needing to work with speakers of a variety of languages.

### 28.5.2 Use of MT for WWW and chat rooms

One of the most important developments in the world of MT in recent years has been its ready availability both in the form of inexpensive software on sale in computer stores (or, in Japan at least, provided free as part of the operating system) and also, famously, on the web. First installed experimentally by CompuServe, *web-page translation*—in the form of the Systran system—is available via the AltaVista search engine at the touch of a button. Users can paste text into a translation window, or give the URL of the web page that they wish to see translated. Several other websites offer free translation using a variety of MT systems. Another recent innovation is *email translation*, and a 'translate' button on some *chat-room* sites.

This increased visibility of MT has had a number of side effects. It has of course increased the general public's awareness of MT, in some cases clarifying its limits but also its benefits. Informal reports suggest that users are at first impressed, then disappointed as they realize its limitations, but finally pragmatic as they learn to get the best out of raw MT. One thing to notice is how using MT to translate the unrestricted (and sometimes poorly written) material that is found on web pages goes against the general recommendations for the use of MT that have been made over the last decade or so. There is certainly a need to *educate* the general public about the low quality of raw MT, and, importantly, *why* the quality is so low. Meanwhile, MT systems have to be adapted and improved so that the quality is raised a little. One particularly important way of doing this, and one which is starting to be addressed, is to tackle the problem of *proper-name translation*. Fortunately, great strides have been made in the

neighbouring field of information extraction towards the *named entity recognition task*, as it is termed (see Chapter 30), and there is evidence that similar techniques can be used to improve MT output so that proper names like *Bill Gates* or *Kanzler Kohl* are not translated as *Addition Barrières* or *Chancellor Cabbage*.

### 28.5.3 Tools for users

As well as research on MT itself, a lot of work has been done recently to develop computer-based tools for translators. Many of the most recent developments have been based on the growing availability of large bilingual corpora, i.e. collections of translated texts in machine-readable form.

A first priority with such corpora is to *align* them, i.e. to establish on a segment-by-segment basis (often paragraphs or sentences) which bits of text in one language correspond to which bits of text in the other. This is not always as straightforward as it may seem, especially when the two languages concerned are typologically very different (so that, for example, the notion of 'sentence' is not compatible), or when the translation is particularly 'free'; but for a lot of texts, this initial alignment is quite successful.

The **aligned bilingual corpus** can then be used as a resource on which can be based a number of tools for the translator. One of these, now widely used, is the **translation memory** (TM) already mentioned above. First proposed in the 1970s, the idea is that the translator can consult a database of previous translations, usually on a sentence-by-sentence basis, looking for anything similar enough to the current sentence to be translated, and can then use the retrieved example as a model. The key to the process is efficient storage of the sentences in the TM, and, most importantly, an efficient *matching* scheme. In current commercial TM systems, the matching is essentially based on character-string similarity, but one could envisage a more sophisticated method, incorporating linguistic 'knowledge' of inflection paradigms, synonyms, and even grammatical alternations. To exemplify, consider (28.9*a*). The example (28.9*a*) differs only in a few characters, and would be picked up by any currently available TM matcher. (28.9*c*) is superficially quite dissimilar, but is made up of words which are related to the words in (28.9*b*) as either grammatical alternatives or near synonyms. (28.9*d*) is very similar in meaning to (28.9*a*), but quite different in structure. Arguably, any of (28.9*a–d*) should be picked up by a sophisticated TM matcher.

(28.9) *a*. When the paper tray is empty, remove it and refill it with paper of the appropriate size.
*b*. When the tray is empty, remove it and fill it with the appropriate paper.
*c*. When the bulb remains unlit, remove it and replace it with a new bulb
*d*. You have to remove the paper tray in order to refill it when it is empty.

As mentioned above, current TMs make no attempt to identify which parts of the

translation correspond to the matched elements of the example: that is up to the translator to decide. For example, if (28.10) is the sentence to be translated, and the TM contains (28.11*a*) with its accompanying translation (28.11*b*), the TM may be able to highlight the differences between (28.10) and (28.11*a*), as we do here, but it is unable to identify which words in (28.11*b*) have to be changed.

(28.10) The large paper tray can hold up to four hundred sheets of A4 paper.

(28.11) *a.* The small paper tray can hold up to three hundred sheets of A5 paper.
*b. Die kleine Papierkassette fasst bis zu dreihundert Blatt in A5-Format.*

Another useful tool based on aligned bilingual corpora is a **bilingual concordancer**. A concordancer in general is a software tool that allows the user to see how a word or phrase is used throughout a text. Sometimes called a *KWIC-index* (keyword in context), it is a tool that literature scholars have used for many years: for example, Fig. 28.1 shows a concordance of the word *curious* in Lewis Carroll's *Alice's Adventures in Wonderland*. A bilingual concordance gives the same sort of listing, but each line is linked to the corresponding translation. This enables the translator to see how a particular word—or more usefully a phrase or a technical term—has been translated before.

Another tool that has been discussed amongst researchers, but has not yet been developed commercially, might be called a *translator-friendly word-processor*. Here is

| | | |
|---|---|---|
| 1 | hed it off. * * * 'What a | curious feeling!' said Alice; 'I must b |
| 1 | against herself, for this | curious child was very fond of pretendi |
| 2 | | 'Curiouser and curiouser!' cried Alice ( |
| 2 | 'Curiouser and | curiouser!' cried Alice (she was so muc |
| 2 | Eaglet, and several other | curious creatures. Alice led the way, |
| 4 | -- and yet—it's rather | curious, you know, this sort of life! |
| 6 | eir heads. She felt very | curious to know what it was all about, |
| 6 | out a cat! It's the most | curious thing I ever saw in my life!' S |
| 7 | ht into it. 'That's very | curious!' she thought. 'But everything' |
| 7 | hought. 'But everything's | curious today. I think I may as well g |
| 8 | Alice thought this a very | curious thing, and she went nearer to w |
| 8 | she had never seen such a | curious croquet-ground in her life; it |
| 8 | seen, when she noticed a | curious appearance in the air: it puzz |
| 9 | next, and so on.' 'What a | curious plan!' exclaimed Alice. 'That's |
| 10 | : 'and I do so like that | curious song about the whiting!' 'Oh, |
| 10 | th, and said 'That's very | curious.' 'It's all about as curious a |
| 10 | ous.' 'It's all about as | curious as it can be,' said the Gryphon |
| 11 | moment Alice felt a very | curious sensation, which puzzled her a |
| 11 | er the list, feeling very | curious to see what the next witness wo |
| 12 | ad!' 'Oh, I've had such a | curious dream!' said Alice, and she tol |
| 12 | her, and said, 'It *was* a | curious dream, dear, certainly: but no |

**Fig. 28.1 Concordance of the word *curious* in *Alice's Adventures in Wonderland***

envisaged software with the normal word-processing facilities enhanced to facilitate the sort of text editing 'moves' that a translator (or, perhaps, a translator working as a posteditor on some MT output) commonly makes. Simple things like transposing two words at the touch of a function key are easy to imagine, but the software could incorporate more linguistically sophisticated tools such as *grammar-conscious global replace*. Imagine you had a text in which the word *fog* had been translated as *brouillard* in French, but you decided *brume* was a better translation. Globally changing *brouillard* to *brume* is only half the job: *brouillard* is masculine, while *brume* is feminine, so some other changes (gender of adjectives and pronouns) may have to be made. A linguistically sophisticated word-processor could do it for you. Similarly, if you wanted to change *buy* to *purchase*, it would be nice if it could automatically change *buying* to *purchasing*, *bought* to *purchased*, etc. The translator-friendly word-processor could also search for 'false friends' (e.g. *librairie* as a translation of *library*) and other 'interference' errors, if the user is a competent but not fluent writer of the target language. Coming back to the idea of parallel text alignment, a similar tool could check the source and target texts to see if any of the source text had inadvertently been omitted in the translation. And the bilingual concordance tool can be used on the current translation texts to check for consistency of translation, e.g. of technical terms.

## Further reading and and relevant resources

The field of MT is a fast-moving one. Latest research and developments are reported in the field's premier journal, *Machine Translation*, and at its conferences, the MT Summit, organized biannually by one of the three regional organizations which make up IAMT, the International Association for Machine Translation (namely AMTA, the Association for MT in the Americas; EAMT, the European Association for MT; and AAMT, the Asian Association for MT).

A comprehensive review of EBMT techniques appeared as Somers (1999). The IBM statistical MT system is described in Brown et al. (1990); a general overview of the approach appears in Knight (1997). A good example of a multi-engine system is PANGLOSS (Frederking et al. 1994)

Anaphora resolution in MT is the subject of a special issue of *Machine Translation* (Mitkov 1999).

KBMT research at Carnegie-Mellon is described in Nirenburg et al. (1992). Latest ideas on interlingua-based MT have been presented at a series of AMTA SIG-IL Workshops, the latest of which was at the MT Summit in Santiago de Compostela (Farwell and Helmreich 2001).

Spoken language MT is represented by major research programmes such as VERBMOBIL (Wahlster 2000), SLT (Rayner et al. 2000), the C-Star consortium of several projects and not a few others. A collection of articles on SLT is assembled in Krauwer (2000).

MT and other resources for minority languages was the subject of a workshop at LREC (Ó Cróinín 2000), which includes a discussion by the present author of the particular case of NIMLs. Frederking et al. (2000) discuss DIPLOMAT, an example of a quickly developed ('rapid ramp-up') interactive MT system for Haitian Creole. Jones and Havrilla (1998) and Nirenburg and Raskin (1998) propose general approaches to this problem.

Possible uses of low-quality MT output were first discussed by Church and Hovy (1993). Systran's web-based MT system is described by Yang and Lange (forthcoming). Flournoy and Callison-Burch (2000) describe Amikai's chat-room translation system. Proper-name recognition is discussed by Turcato et al. (2000) in connection with their TV closed-caption translation system.

Translation tools are the subject of Isabelle and Church (1998). Research involving aligned parallel bilingual texts has been copious in recent years. Overviews of the main issues are to be found in Manning and Schütze (1999:463 ff.); Véronis (2000); Wu (2000); Melamed (2001), among others. Hutchins (1998) gives a historical perspective of translators' tools, including the translation memory. Researchers at RALI in Montreal have developed a number of corpus-based translators' tools including the bilingual concordancer (Macklovitch, Simard, and Langlais 2000) and some of the other tools mentioned.

## References

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. 'A statistical approach to machine translation'. *Computational Linguistics*, 16(2), 79–85.

Church, K. W., and E. H. Hovy. 1993. 'Good applications for crummy machine translation'. *Machine Translation*, 8(4), 239–58.

Farwell, D., and S. Helmreich (eds.). 2001. *Proceedings of the 5th Workshop on Interlinguas and Interlingual Approaches to MT*, Summit VIII. Santiago de Compostela, Spain.

Flournoy, R. S., and C. Callison-Burch. 2000. 'Reconciling user expectations and translation technology to create a useful real-world application'. *Translating and the Computer 22: Proceedings of the 22nd International Conference on Translating and the Computer*. London.

Frederking, R., S. Nirenburg, D. Farwell, S. Helmreich, E. Hovy, K. Knight, S. Beale, C. Domashnev, D. Attardo, D. Grannes, and R. Brown. 1994. 'Integrating translations from multiple sources within the Pangloss Mark III machine translation system'. *Technology Partnerships for Crossing the Language Barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, 73–80. Columbia, Md, USA.

——A. Rudnicky, C. Hogan, and K. Lenzo. 2000. 'Interactive speech translation in the DIPLOMAT Project'. *Machine Translation*, 15(1–2).

Hutchins, J., 1998. 'The origins of the translator's workstation'. *Machine Translation*, 13(4), 287–307.

Isabelle, P., and K. W. Church (eds.). 1998. 'New tools for human translators', special issue of *Machine Translation*, 13(1–2).

Jones, D., and R. Havrilla. 1998. 'Twisted pair grammar: support for rapid development of machine translation for low density languages'. In D. Farwell, L. Gerber, and E. Hovy (eds.), *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas AMTA '98*. Berlin: Springer, 318–32.

Knight, K. 1997. 'Automating knowledge acquisition for machine translation'. *AI Magazine*, 18(1), 81–96.

Krauwer, S. (ed.). 2000. 'Spoken language translation', special issue of *Machine Translation*, 15(1–2).

Macklovitch E., M. Simard M., and P. Langlais. 2000. 'TransSearch: a Free translation memory on the world wide web'. *Second International Conference on Language Resources and Evaluation* (*LREC*), 1201–8, Athens, Greece.

Manning, C. D., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Melamed, I. D. 2001. *Empirical Methods for Exploiting Parallel Texts*. Cambridge, Mass.: MIT Press.

Mitkov, R. (ed.). 1999. 'Anaphora resolution in machine translation', special issue of *Machine Translation*, 14(3–4).

Nagao, M. 1984. 'A framework of a mechanical translation between Japanese and English by analogy principle'. In A. Elithorn and R. Banerji (eds.), *Artificial and Human Intelligence*. Amsterdam: North-Holland, 173–80.

Nirenburg, S., and V. Raskin, 1998. 'Universal grammar and lexis for quick ramp-up of MT systems', *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (*COLING-ACL '98*), 975–9. Montreal, Canada.

——J. Carbonell, M. Tomita, and K. Goodman. 1992. *Machine Translation: A Knowledge-Based Approach*. San Mateo, Calif.: Morgan Kaufmann.

Ó Cróinín, D. (ed.). 2000. *Proccedings of LREC 2000: The 2nd International Conference on Language Resources and Evaluation Workshop, Developing Language Resources for Minority Languages: Reusability and Strategic Priorities* (Athens).

Rayne, M., D. Carter, P. Bouillon, V. Digalakis, and M. Wirén. 2000. *The Spoken Language Translator*. Cambridge: Cambridge University Press.

Sato, S., and M. Nagao, 1990. 'Toward memory-based translation'. *COLING-90: Papers Presented to the 13th International Conference on Computational Linguistics*, iii, 247–52. Helsinki, Finland.

Somers, H. 1999. 'Review article: example-based machine translation'. *Machine Translation*, 14(2), 113–57.

Turcato, D., F. Popowich, P. McFetridge, D. Nicholson, and J. Toole 2000. 'Pre-processing closed captions for machine translation'. *Proceedings of ANLP/NAACL 2000 Workshop: Embedded Machine Translation Systems*, 38–45. Seattle, USA.

Véronis, J. (ed.). 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Press.

Wahlster, W. (ed.). 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer.

Wu, D. 2000. 'Alignment'. In R. Dale, H. Moisl, and H. Somers (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, Inc., 415–58.

Yang, J., and E. Lange. 2002. 'Going live on the Internet'. In H. Somers (ed.), *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins.