[Article intended for *Erhverv og Sprog*, 1995, but not published]

TRENDS IN MACHINE TRANSLATION RESEARCH

John Hutchins

(University of East Anglia, Norwich, England)

Since the beginning of the 1990s the methods used in research on machine translation have undergone a marked change. Although it will be some time before MT systems on the market show any signs of benefiting from these new approaches, translators and others interested in MT in general should be aware of these developments.

Rule-based approaches

During the 1980s the dominant framework of MT research was the approach based essentially on linguistic rules of various kinds: rules for syntactic analysis, lexical rules, rules for lexical transfer, rules for syntactic generation, rules for morphology, etc. Although the so-called `transfer' systems dominated the scene (e.g. Ariane, Metal, SUSY, and Eurotra), there were also various `interlingua' systems. Some were still essentially linguistics-oriented (DLT and Rosetta), but others adopted knowledge-based approaches, making use of non-linguistic information related to the domains of texts to be translated.

Research on all these lines has continued into the 1990s. The transfer approach is found in the continuing research on the Metal system, the MT research at SITE on developments of Eurolang, and the recently marketed LMT system from IBM. The interlingua approach is to be found in the knowledge-based approach at Carnegie Mellon University (in particular, the large-scale project for Caterpillar should be noted), in the ULTRA system at the New Mexico State University, in the UNITRAN system based on the linguistic theory of Principles and Parameters, and in the joint Pangloss project at three US centres.

A characteristic feature of rule-based systems is the transformation or mapping of labelled tree representations, e.g. (as in Eurotra) from a morphological tree into a syntactic tree, from a syntactic tree into a semantic tree, from an interface tree of the source language into an equivalent target-language tree, and so forth. Transduction rules require the satisfaction of precise conditions: a tree must have a specific structure and contain particular lexical items or specific syntactic or semantic features. In addition, every tree is tested by formation rules; i.e. a `grammar' confirms the acceptability of its structure and the relationships it represents. Grammars and transduction rules specify the `constraints' which determine the possibility of transfer from one level to another and ultimately the transfer of a source-language text to a target-language text.

Since the mid 1980s there has been a general trend towards the adoption of `unification' and `constraint-based' formalisms. Their main advantage is the simplification of the rules (and hence the computational processes) of analysis, transformation and generation. Instead of a series of complex multi-level representations and large sets of rules, which apply only in very specific circumstances and to specific representations, there are monostratal representations and a restricted set of abstract rules, with conditions and constraints incorporated into specific lexical

entries. At the same time, the components of these grammars are in principle reversible, so that it is no longer necessary to construct for the same language different grammars of analysis and generation.

Lexical information

The syntactic orientation which characterised transfer systems in the past has thus been replaced by `lexicalist' approaches, with a consequential increase in the range of information attached to lexical units the lexicon: not just morphological and grammatical data and translation equivalents, but also information on syntactic and semantic constraints and non-linguistic and conceptual information. The expansion of lexical data is seen most clearly in the lexicons of interlingua-based systems, which include large amounts of non-linguistic information.

The construction of MT lexicons is a complex and expensive task and many research groups are investigating methods of extracting lexical information from readily available lexicographic sources, such as bilingual dictionaries intended for language learners, specialised technical dictionaries, and the terminological databanks used by professional translators. At the same time, there is much more collaboration in the construction of lexicons for a wide range of natural language applications, not just for machine translation but also for text analysis and information retrieval (e.g. the Electronic Dictionary Research project supported by several Japanese computer manufacturing companies.)

Corpus-based methods

Since 1989 the dominance of the rule-based framework has been broken by the emergence of new methods and strategies collectively known as `corpus-based' methods, and it is these developments, above all, which justify the view that MT has entered a new era.

In 1989 a group from IBM published the results of experiments on a system based purely on a statistical approach. The distinctive feature of the IBM Candide system is that statistical methods are used as virtually the sole means of analysis and generation; no linguistic rules are applied. The IBM research was undertaken on the vast corpus of French and English texts contained in the reports of Canadian parliamentary debates (the Canadian Hansard). The essence of the IBM method is first to align phrases, word groups and individual words of the parallel bilingual 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.

What surprised most researchers (particularly those involved in rule-based approaches) was 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. The IBM researchers have sought to improve these results, and propose to introduce more sophisticated statistical methods. Rather surprisingly, they also intend to make use of some linguistic rules of morphology and syntax, even though apparently they had set out to demonstrate the inadequacies of traditional rule-based approaches.

The second major `corpus-based' approach - benefiting likewise from improved rapid access to large databanks of text corpora - is what is known as the `example-based' (or `memory-based') approach. The underlying notion is that translation often involves the finding or recalling of analogous examples, the discovery or recollection of how a particular expression or some similar

phrase has been translated before. The example-based approach is founded on processes of extracting and selecting equivalent phrases or word groups from a databank of parallel bilingual texts, which have been aligned either by statistical methods (similar perhaps to those used by the IBM group) or by more traditional `rule-based' methods of analysis. For calculating matches, some MT groups use semantic methods, e.g. a semantic network or a hierarchy (thesaurus) of domain terms. Other groups use statistical information about lexical frequencies in the target language. The main advantage of the approach is that since the texts have been extracted from databanks of actual translations previously produced by professional translators there is an assurance that the results should be accurate and idiomatic.

Hybrid systems

The general view of many experts is that future MT systems will combine these newer corpus-based methods with traditional rule-based approaches, i.e. they will be `hybrid' systems. For example, the linguistic methods of the traditional systems might provide the foundation upon which processes involving domain-specific knowledge banks, statistical data and examples of translated texts will operate. In this approach, the linguistic rules will be somewhat less ambitious and complex than those of current systems, e.g. syntactic analysis may well be limited to the recognition of surface phrase structures and dependencies, semantic analysis will be more limited, and lexical information will be extracted mainly from standard sources such as general-purpose dictionaries. Corpus-based methods would then be used to refine the rule-based analyses, to improve lexical selection and to generate more idiomatic target language texts. It should be stressed, however, that the newer corpus-based approaches have yet to be fully tested in experimental systems, and it is unlikely that any commercial system using these methods is to be expected before the end of the century.

Most future systems will be directly integrated in general computer-based systems for the production, transmission and management of documents, and will contribute therefore to more sophisticated workbenches for translators. Indeed this is happening already. A by-product of the `corpus-based' approach is the method of aligning bilingual parallel texts to provide means of accessing `translation memories' of previously translated texts, and such facilities are already commercially available in the workstations from STAR and TRADOS.

Special-purpose systems

Until the mid 1980s it was a general assumption that MT systems should be capable in principle of dealing with the full range of written language, i.e. that they should be general-purpose systems. In practice, however, systems are limited to particular ranges of subjects, since large dictionaries are needed and developers have concentrated on domains where there is greatest demand.

It has long been recognised that concentration on a sublanguage (e.g. weather reports in the well-known Meteo system from the mid 1970s) eases considerably the difficulties of analysing complex sentences, of selecting correct target language equivalents and of generating idiomatic output. Equally effective have been implementations in environments where the language of input is `controlled' in some respect (e.g. in the Xerox implementation of Systran and in the many successful systems developed by the Smart Corporation.)

During the last decade, more attention has been paid to the design of special-purpose systems. There has continued to be much interest in sublanguage and in controlled language systems, but a

relatively recent development has been the design of systems for (and often by) specific users. These systems are typically restricted in vocabulary and subject domains (sublanguage), and also involve control of input language. For example, Volmac Lingware Services has produced MT systems for a textile company, an insurance company, and for translating aircraft maintenance manuals; Cap Gemini Innovation developed TRADEX to translate military telex messages for the French Army; and in Japan, CSK developed its own ARGO system for translation in the area of finance and economics, and now offers it also to outside clients. Such user-designed systems are an encouraging sign that the computational methods of MT and NLP are now spreading outside the limited circles of researchers. It is a trend which could well expand rapidly in coming years.

New directions

Some of the most ambitious projects at present are those involving spoken language translation. These systems are inevitably highly restricted in domain and range. The Japanese ATR project, underway already for seven years and set to continue into the next century, is a system for registration by telephone at international conferences and for hotel booking by telephone. The German Verbmobil project aims to develop a transportable aid for face to face English-language commercial negotiations by Germans and Japanese who do not know English fluently. The JANUS project - a collaboration involving ATR, Carnegie Mellon and Karlsruhe University - is also restricted to conference registration negotiations. Each group is developing speech recognition and speech synthesis modules for their own languages (Japanese, English, German) and the translation programs linking their language to the other two. A successful public demonstration of an early prototype of JANUS was given in January 1993.

A further feature of the last five years is the recognition of a demand for types of translations which have not previously been considered. In the past, systems were built generally for bilingual users, for translators and for those knowing both source and target languages. The needs of those not knowing the target language were neglected, e.g. businessmen engaged in foreign trade needing to communicate fairly simple standard messages in an unknown language (e.g. confirmation of an order, booking of accommodation, etc.) In recent years, there have been experiments on `dialogue-based MT' where the text to be translated is composed in a collaborative process between man and machine, i.e. another approach to `control' of input. In this way it is possible to construct a text which the system is known to be capable of translating without further reference to the author, which needs no revision, and for which good quality output can be assured.

Current and future demands

Fully automatic systems capable of producing idiomatic texts comparable to human translation are no longer the goal of MT research. It is now largely focused on the development of systems limited to sublanguages or to specific technical fields. For the individual professional translator, it is recognised that full MT as such is not appropriate, hence the concentration on more sophisticated facilities in translator workstations, which provide access to dictionary and terminological data and to previous translations ('translation memories') and where MT software is one option for producing draft versions if required.

The new research developments are taking place against a background of a rapidly expanding marketplace for MT and increasing numbers of users. In recent years, the number of pages translated automatically has increased considerably - at present, more than a million pages annually, or about 300 million words a year. The expansion has taken place primarily in large

multinational companies and in translation agencies, particularly for the translation of technical manuals. It is now well established that in favourable conditions, limited-domain systems which are far from perfect are being used successfully and cost-effectively.

There has, however, also been an increase in the numbers of non-professional users. Many have purchased cheap PC-based systems. These are certainly crude in linguistic terms, and the effectiveness and quality of the systems may be doubtful; but there is no doubt that they are "satisfying" the needs of the users. We can expect an expansion of users of such personal computer systems, and we can also predict an expanding use of MT systems over electronic networks. New systems are likely to appear in the future which meet more closely the specific needs of a wider variety of potential users. The great majority of these MT users will not be from the translation or language professions, but they will be people with relatively little knowledge of foreign languages who just want to find out what a text or document is about or who simply want to communicate with others in another language.