CHAPTER 27

MACHINE TRANSLATION: GENERAL OVERVIEW

JOHN HUTCHINS

Abstract

This chapter introduces the main concepts and methods used for machine translation systems from the beginnings of research in the 1950s until about 1990; it covers the main approaches of rule-based systems (direct, interlingua, transfer, knowledge based), and the principal translation tools; and it concludes with a brief historical sketch. (For methods since 1990 see Chapter 28.)

27.1 Introduction

The term **machine translation** (MT) refers to computerized systems responsible for the production of translations with or without human assistance. A distinction is commonly made between human-aided MT (HAMT) and machine-aided human translation (MAHT). The latter comprises computer-based **translation tools** which support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, stores of previously translated texts ('translation memories'), and integrated resources, commonly referred to as **translator workstations** or translator workbenches. The term computer-aided translation (CAT) is sometimes used to cover all these computer-based translation systems.

In this chapter, we will cover first the main methods and problems of machine translation, then in more detail the main features of the rule-based approaches used until the late 1980s. After outlining the various types of translation aids developed up to 1990, we end with a brief survey of the historical development of MT and translation tools.

27.2 Principal Approaches and Methodologies

Although the ideal goal of MT systems may be to produce high-quality translation without human intervention at any stage, in practice this is not possible except in highly constrained situations (see below). In particular, if the reason for using an MT system is to produce translations of publishable quality, approximating what might be expected from a human translator, then the output must be revised (or, as it is known in MT circles, **postedited**). It should be noted that in this respect MT does not differ from the output of most human translators, which is normally revised by a second translator before dissemination. However, the types of errors produced by MT systems do differ from those of human translators (incorrect prepositions, articles, pronouns, verb tenses, etc.). If the reason for using an MT system is simply to acquire some knowledge of the content of the original text, then the output may be left unedited or only lightly revised. This is frequently the case when the output is intended only for specialists familiar with the text subject. As a further example, unedited output might serve as a rough draft for a human translator, i.e. as a 'pre-translation', or for someone with some familiarity with the target language who needs a working basis for producing an original text.

The translation quality of MT systems may be improved by adjusting (editing or controlling) the input. One option is for input texts to be marked (**pre-edited**) to indicate prefixes, suffixes, word divisions, phrase and clause boundaries, or grammatical categories (e.g. to distinguish the noun *cónvict* and its homonymous verb *convíct*). More common (particularly in large installations) is the **control** of the vocabulary and of the grammatical structures of texts submitted for translation. In this way, the problems of ambiguity and of the selection of equivalents are reduced, in some cases eliminated. Although the costs of preliminary editing may be high, postediting is reduced considerably. Another option is for systems to be specifically designed for particular subject areas (**sublanguages**) or for the needs of specific users. In each case, they reduce the known deficiencies of full-scale general-purpose MT systems. (See Chapter 23 on sublanguages and controlled languages.)

MT systems can be designed either specifically for two particular languages, e.g. Russian and English (**bilingual** systems), or for more than a single pair of languages (**multilingual** systems). Bilingual systems may be designed to operate either in only one direction (unidirectional), e.g. from one **source language** (SL) into one **target language** (TL) only, or in both directions (bidirectional). Multilingual systems are usually designed to be bidirectional; but most bilingual systems are unidirectional.

In overall system design, there have been three basic types. The first (and historically oldest) type is generally referred to as the **direct translation** (or 'binary translation') approach: the MT system is designed in all details specifically for one particular pair of languages. Translation is direct from the source text to the target text, with as little syntactic or semantic analysis as necessary. The basic assumption is that the vocabulary and syntax of SL texts should not be analysed any more than strictly necessary for the resolution of SL ambiguities, for the correct identification of TL expressions, and for the specification of TL word order. In other words, SL analysis is oriented specifically to the production of representations uniquely appropriate for one particular TL. Typically, such systems consist of a single large bilingual dictionary and a single program for analysing and generating texts; such 'direct translation' systems are necessarily bilingual and normally unidirectional.

The second basic design strategy is the **interlingua** approach, which assumes that it is possible to convert SL texts into semantico-syntactic representations which are common to more than one language (but not necessarily 'universal' in any sense). From such interlingual representations texts are generated into other languages. Translation is thus in two stages: from SL to the interlingua and from the interlingua to the TL. Procedures for SL analysis are intended to be SL specific and not oriented to any particular TL; likewise programs for TL synthesis are TL specific and not designed for input from particular SLs. A common argument for the interlingua approach is economy of effort in a multilingual environment, i.e. an analysis program for a particular SL can be used for more than one TL, and a generation program for a particular TL can be used again. On the other hand, the complexity of the interlingua itself is greatly increased. Interlinguas may be based on a 'logical' artificial language, on a 'natural' auxiliary language (such as Esperanto), a set of semantic primitives common to all languages, or a supposedly 'universal' vocabulary.

The third basic strategy is the less ambitious **transfer** approach. Rather than operating in two stages via a single interlingual representation, there are three stages involving underlying (abstract) representations for both SL and TL texts. The first stage converts SL texts into abstract SL-oriented representations; the second stage converts these into equivalent TL-oriented representations; and the third generates the final TL texts. Whereas the interlingua approach necessarily requires complete resolution of all ambiguities in the SL text so that translation into any other language is possible, in the transfer approach only those ambiguities inherent in the language itself are tackled (e.g. homonyms and ambiguous syntactic structures). Problems of lexical differences between languages are dealt with in the second stage (transfer proper). 'Transfer' systems consist therefore typically of three types of dictionaries: SL dictionaries containing detailed morphological, grammatical, and semantic information; similar TL dictionaries; and a bilingual 'transfer' dictionary relating base SL forms and base TL forms. Likewise, 'transfer' systems have separate grammars for SL analysis, TL synthesis, and for the transformation of SL structures into equivalent TL forms.

Within the stages of analysis and synthesis, most MT systems exhibit clearly separated components involving different levels of linguistic description: morphology, syntax, and semantics. Hence, **analysis** may be divided into morphological analysis (identification of word endings, word compounds), syntactic analysis (identification of phrase structures, dependency, subordination, etc.), and semantic analysis (resolution of lexical and structural ambiguities); and **synthesis** (or **generation**) may likewise pass through stages of semantic synthesis (selection of appropriate compatible lexical and structural forms), syntactic synthesis (generation of required phrase and sentence structures), and morphological synthesis (generation of correct word forms). In 'transfer' systems, the **transfer** component may also have separate programs dealing with lexical transfer (selection of vocabulary equivalents) and with structural transfer (transformation into TL-appropriate structures). In some earlier forms of transfer systems, analysis did not involve a semantic stage; transfer was restricted to the conversion of syntactic structures, i.e. 'syntactic transfer'.

In many older systems, particularly those of the 'direct translation' type, the components of analysis, transfer, and synthesis were not always clearly separated. In some there was also a mixture of data (dictionary and grammar) and processing rules and routines. Later systems exhibited various degrees of **modularity**, so that system components, data, and programs can be adapted and changed with minimal damage to overall efficiency. A further development in more recent systems is the design of representations and processing rules for **reversibility**, i.e. the data and transformations used in the analysis of a particular language are applied in reverse when generating texts in that language.

The direct translation approach was typical of the 'first generation' of MT systems (from the 1950s to the mid-1970s). The interlingua and transfer-based systems were characteristic of the 'second generation' system (during the 1970s and 1980s). Both are based essentially on the specification of rules (for morphology, syntax, lexical selection, semantic analysis, and generation). They are thus now known generically as **rule-based systems**, distinguishing them from more recent corpus-based approaches (Chapter 28).

27.3 Rule-Based Systems

The major problem for translation systems, whatever the particular strategy and methodology, concerns the resolution of lexical and structural ambiguities, both within languages (monolingual ambiguity) and between languages (bilingual ambiguity). The problems of monolingual disambiguation are not, of course, unique to translation; they are present in all natural language processing and are dealt with elsewhere in this volume.

Any monolingual ambiguity is a potential difficulty in translation since there may be more than one possible TL equivalent. Homographs and polysemes (English *cry*, French *voler*) must be resolved before translation (French *pleurer* or *crier*, English *fly* or *steal*); ambiguities of grammatical category (English *light* as noun, adjective, or verb, *face* as noun or verb) must likewise be resolved for choice between French *lumière*, *clair*, or *allumer*, and between *visage* or *confronter*, etc. Examples of monolingual structural ambiguities occur when a word or phrase can potentially modify more than one element of a sentence. In *old men and women*, the adjective *old* may refer only to *men* or to both *men* and *women* (French *vieux et femmes* or *vieux et vieilles*). In English, prepositional phrases can modify almost any preceding verb or noun phrase, e.g. (*a*) *The car was driven by the teacher at high speed*, (*b*) *The car was driven by the teacher with defective tyres*, and (*c*) *The car was driven by the teacher with red hair*. Lexical and structural ambiguities may and often do occur together: *He saw her shaking hands*, where *shaking* can be either an adjective ('hands which were shaking') or a gerundive verb ('that she was shaking hands').

Bilingual lexical ambiguities occur primarily when the TL has distinctions absent in the SL: English *river* can be French *rivière* or *fleuve*, German *Fluss* or *Strom*; English *eat* can be German *essen* or *fressen*; English *wall* can be French *mur* or *paroi*, German *Wand*, *Mauer*, or *Wall*. Even the apparently simple adjective *blue* can be problematic: in Russian a choice must be made between *sinii* (dark blue) and *goluboi* (light blue). A more extreme, but not uncommon, example is illustrated by the translation of the verb *wear* from English to Japanese; although there is a generic verb *kiru* it is normal to use the verb appropriate to the type of item worn: *haoru* (coat or jacket), *haku* (shoes or trousers), *kaburu* (hat), *hameru* (ring or gloves), *shimeru* (belt, tie, or scarf), *tsukeru* (brooch or clip), and *kakeru* (glasses or necklace).

Such choices can be circumvented by TL 'cover' words, i.e. selecting a single, most generally acceptable TL equivalent for a SL homograph, or by including phrases as dictionary items, e.g. for idioms (*wage war*), for compounds with specific TL equivalents which cannot be derived from components (*make away with* and *faire disparaître*, *look up* and *aufsuchen*), as well as for technical terms which have standardized translations (*plug connector* as *raccord de fiche*).

Bilingual structural differences can be either general (e.g. in English adjectives generally precede nouns but in French they usually follow); or they can be specific to particular structures, e.g. when translating the English verb *like* (*She likes to play tennis*) as a German adverb *gern* (*Sie spielt gern Tennis*); or they can be determined by particular lexical choices, e.g. an English simple verb (*trust*) rendered by a French circumlocution (*avoir confiance à*). Not uncommonly, structural differences combine with lexical differences, e.g. the translation of *know* into French or German, where choice of *connaître* (*kennen*) or *savoir* (*wissen*) affects both structure (*Je connais l'homme*, *Ich kenne den Mann: Je sais ce qu'il s'appelle*, *Ich weiss wie er heisst*) and the translation of other lexical items (*what* as *ce que* and *wie*).

The tools available are familiar from other fields of computational linguistics: the provision of dictionaries with lexical, grammatical, and translational information; the use of morphological and syntactic analysis to resolve monolingual ambiguities and to derive structural representations; the use of contextual information, of semantic features, of case markers, and of non-linguistic ('real world') information to resolve semantic ambiguities. The information required for resolution may be applied at any stage, during analysis of the SL text, during generation of the TL text, or at a transfer stage.

Dictionaries contain the information necessary for SL analysis (morphological variants, syntactic functions, semantic features, etc.) and for TL synthesis (translation equivalents, constraints on TL syntax and word formation, etc.). There may be a single bilingual dictionary, as in many older 'direct' systems, or, more commonly, there may be separate dictionaries for analysis (monolingual SL dictionary), transfer (bilingual SL–TL dictionary), and synthesis (monolingual TL dictionary). Dictionaries may contain entries either in full forms or in only base ('canonical' or root) uninflected forms, if these can be readily identified from inflected forms. In general, irregular forms are entered in full.

Morphological analysis is concerned with the identification of base forms from inflected forms, both regular (*fake*:*faked*) and irregular (*make*:*made*). It may also involve the recognition of derivational forms (e.g. English *-ly* as an adverb derived from an adjective, German *-heit* as a noun from an adjective). All MT systems have problems with 'unknown' words, primarily neologisms (common in scientific and technical literature), but also unanticipated combinations. If derivational elements can be correctly identified then some attempt can be made to translate, particularly in the case of 'international' prefixes and suffixes (e.g. French *demi-* and English *semi-*, French *-ique* and English *-ic*). Morphological analysis often includes the segmentation of compounds, e.g. in German. However, segmentation can be problematic, e.g. *extradition* might be analysed as either *extradit+ion* or *ex+tradition*, *cooperate* as either *co+operate* or *cooper+ate*. These would be resolvable by dictionary consultation, but sometimes alternative segmentations are equally valid (German *Wachtraum* could be *guard room* (*Wacht+Raum*) or *day dream* (*Wach+Traum*). (For more on morphological analysis see Chapter 2.)

As in other areas of computational linguistics, there have been three basic approaches to syntactic structure analysis. The first aims to identify legitimate sequences of grammatical categories, e.g. English article, adjective, noun. This approach led to the development of parsers based on *predictive analysis*, where a sequence of categories enables the prediction of a following category. The second approach aims to recognize groups of categories, e.g. as noun phrases, verb phrases, clauses, and ultimately sentences. These parsers are based on *phrase structure* or constituency grammar. The third approach aims to identify dependencies among categories, e.g. reflecting the fact that prepositions determine the case forms of German and Russian nouns. The basis for these parsers is *dependency grammar*. Each approach has its strengths and weaknesses, and systems often adopt an eclectic mixture of parsing techniques (Chapter 12).

SL structures are transformed into equivalent TL structures by conversion rules, in the case of phrase structure or dependency trees by *tree transducers*, which may apply either unconditionally (e.g. English adjective+noun to French noun+adjective) or conditionally, triggered by specific lexical items (e.g. English *like* to German *gern*). Structural synthesis of TL sentences is similar: some syntax and morphology rules apply unconditionally (e.g. formation of English passives, case endings of German nouns after particular prepositions), others are conditional (irregular verb forms).

Semantic analysis concerns the resolution of problems remaining after morphological and syntactic analysis. While the latter can resolve problems of category ambiguity (e.g. whether a particular occurrence of *light* is a noun, a verb, or an adjective), semantic analysis must decide whether the homograph adjective *light* is being used to mean 'not heavy' or 'not dark'. Likewise semantic analysis is needed to resolve structural ambiguity (e.g. the *shaking hands* example), and any bilingual lexical differences (such as the *know* and *wear* examples). Two basic means are commonly employed. The first is the use of semantic features attached to dictionary items, e.g. the two senses of French *voler* may be distinguished by semantic features to indicate that in its 'flying' sense its subject (grammatical or logical) can be a 'bird' or a 'plane' and in its 'stealing' sense it may be a 'man'. Problems of structural ambiguity can also be resolved using semantic features: e.g. to avoid the mistranslation of *pregnant women and children* into French *femmes and enfants enceintes*, the features for *pregnant* might restrict its use to the modification of 'female' nouns and might exclude its attachment to 'young' nouns.

The second approach is through the identification of thematic (or 'deep' case) roles such as the agents, recipients, instruments, and locations of actions. Although languages differ in the expression of cases (English and French via prepositions and word order, German and Russian via grammatical noun case endings, Japanese via particles, etc.) and few surface markers are unambiguous (English *with* may express manner, attribute, or instrument), there is sufficient universality of underlying meanings and structures to encourage their widespread use in MT systems.

Semantic features and case roles may be adopted as **universals** in interlingua and transfer systems. Further steps towards interlingual representations have included the decomposition of lexical items into semantic 'primitives' (a basic set of components sufficient to distinguish meanings) and the analysis of structures into logical forms, e.g. in terms of predicates and arguments. A major difficulty with such analyses is the loss of surface information which may be essential to generate appropriate TL sentences: a logical analysis may disregard theme and rheme structure and may ignore differences of active and passive formation. The main problem for interlingua systems, however, is the treatment of bilingual lexical differences and specifically whether the interlingua should reflect every semantic difference in all languages involved, e.g. the Japanese *wear* distinctions even when translating between French and English.

A number of problems resist traditional linguistic treatment. The identification of the antecedent of a pronoun may well depend on (non-linguistic) knowledge of events or situations: *The soldiers killed the women. They were buried next day*. We know that the pronoun *they* does not refer to *soldiers* and must refer to *women* because we know that 'killing' implies 'death' and that 'death' is followed (normally) by 'burial'. This identification is crucial when translating into French where the pronoun must be *elles* and not *ils*. (For the treatment of anaphora see Chapter 14.)

Non-linguistic knowledge can be applied to many transfer problems, e.g. whether a *wall* is interior and exterior (*Wand* and *Mauer*), whether a *river* flows into the sea or not (*fleuve* versus *rivière*), whether the object modified is normally dark blue or light blue (*sinii* versus *goluboi*), etc. Such examples and many others are reasons for including **knowledge bases** in MT systems, either as adjuncts to traditional semantic analyses or as the basic mechanisms of lexical analysis and transfer.

Historically, rule-based MT systems have progressively introduced 'deeper' levels of analysis and transfer. Early word-for-word systems were restricted to bilingual dictionaries and simple morphology. Later 'direct' systems introduced syntactic analysis and synthesis. Phrase-structure and dependency analyses provided the basis for simple transfer systems with little semantic analysis ('syntactic transfer'). The addition of semantic features and case relations has led to the now more common type of 'semantico-syntactic' transfer system. The more extensive introduction of interlingual or quasi-universal items and structures characterizes 'advanced' transfer designs and, of course, interlingua systems. Finally, full conceptual-semantic analysis is a feature of **knowledge-based** interlingua systems based on or incorporating various AI techniques.

The characteristic feature of rule-based systems is the transformation ('transduction') of labelled tree representations, e.g. from a morphological tree into a syntactic tree, from a syntactic tree into a semantic tree, etc. Transduction rules require the satisfaction of precise conditions: a tree must have a specific structure and contain particular lexical items or particular (syntactic or semantic) features. In addition, every tree is tested by a 'grammar' for the acceptability of its structure at the level in question: morphological, syntactic, semantic, etc. Grammars and transduction rules specify the 'constraints' determining transfer from one level to another and hence, in the end, the transfer of a source language text to a target language text. Failure at any stage means failure of the whole process.

Since the mid-1980s there has emerged a widely accepted framework, embracing many variants of *unification* and *constraint-based* formalisms (see also Chapters 4 and 12). Instead of the large set of transduction rules devised only for very specific circumstances, these formalisms offer a restricted set of abstract rules and require conditions of application to be incorporated into lexical entries. For example, the translation of a sentence including *likely* (e.g. *The student is likely to work*) into a French equivalent with *probable* (*Il est probable que l'étudiant travaillera*) involves transformation of an English infinitival complement (*to work*) to a French subordinate clause (*que . . . travaillera*); the conditions for effecting the conversion are included in the feature sets of the lexical entries for *likely* and *probable*; no transduction rules specific for these constructions are required.

The main advantage of unification and constraint-based grammars 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 there are mono-stratal representations and/or simple lexical transfer; and the syntactic orientation which characterized previous transfer systems is replaced by lexicalist solutions. In addition, grammars are in principle *reversible*; the same formalism can in theory be applied in both analysis and synthesis.

27.4 Translation Tools

MT systems are not suitable for use by professional translators, who do not like to have to correct the irritatingly 'naive' mistakes made by computer programs. They prefer computer aids that are under their full control. Since the 1960s and 1970s translators have been using programs for accessing approved translations of subject terminology and for managing their own glossaries. An early computer-based aid was the provision of 'text-related glossaries', produced by matching individual words of a particular text against an existing bilingual terminology database. Next came facilities for on-line access to multilingual **termbanks**, and programs for **terminology management** by individual translators, essential with rapidly changing terminology in many scientific and technical fields and with the need for consistency in the translation of company and technical documentation.

The value of easy access to previous translations, either in whole or in part, has long been recognized by translators, particularly when dealing with repetitive texts or with updated versions of already translated documents (a frequent event in large organizations and companies). However, it is only since the end of the 1980s, with the increased availability of large electronic corpora of bilingual texts and of appropriate statistical methods for organizing and accessing them, that the **translation memory** has become an invaluable translation tool, and with it the integration of various translation tools (and MT systems) in the **translator workstation**, for producing good-quality machine-aided translations (see Chapter 28).

27.5 Brief Historical Sketch

The use of computers for translation was first proposed in 1947 by Warren Weaver in correspondence with Norbert Wiener and in conversation with A. D. Booth, and in 1949 in his memorandum sent to American scientists. The next few years saw the beginning of research in the United States, the Soviet Union, and Western Europe. Projects were wide ranging: dictionary (lexicographic) work, 'direct translation' models, syntactic parsing, interlinguas, statistical analyses, dependency grammars, stratificational models, mathematical linguistics, etc.; and led to the beginnings of computational linguistics, artificial intelligence, formal linguistics, and non-numeric programming languages, etc.

However, the practical results were disappointing; by the mid-1960s there were few working translation systems, and the output quality was unsatisfactory. In 1966, a committee set up by US sponsors of research (ALPAC) concluded that there was little prospect of good-quality and/or cost-effective MT, and there were enough human translators to cope. Many projects were ended, not just in the USA, but worldwide. However, in the next ten years there was a gradual revival in Europe and Canada; the turning point in 1976 was the *Météo* 'sublanguage' system (see Chapter 23), and the installation of the Systran system at the European Commission.

Research quickened in Europe and in Japan, a common framework being the multi-level (stratal) 'transfer' approach, e.g. in the Ariane system (Grenoble), the SUSY system (Saarbrücken), the Mu system (Kyoto), and the multi-national Eurotra project of the European Commission. During the 1980s, the 'interlingua' approach was investigated in the Netherlands (Rosetta and DLT) and at Carnegie-Mellon University—particularly notable for its knowledge-based approach.

During the 1980s, MT systems came into practical operation at numerous installations (particularly multinational corporations), invariably involving extensive pre-editing and postediting and/or the use of controlled language input. At the same time, systems for personal computers began to be marketed—Japanese computer companies such as Fujitsu, Hitachi, and NEC were major players.

At the end of the decade came the translator workstation, providing the human translator with facilities that genuinely increased productivity. Terminology control and improved text processing facilities had become familiar during the 1980s, but translation memories were the crucial development.

During the 1990s most research focused on corpora-based methods, and on combining them with traditional rule-based methods; at the cutting edge is the work on spoken language translation. In the marketplace, the situation has been transformed by low-cost (and low-quality) software for personal computers (mostly 'rule based'), by the demand for immediate 'less-than-perfect' translation on the Internet, and by the development of systems for cost-effective large-scale production of company documentation.

Further reading and relevant resources

General introductions to machine translation have been provided by Arnold et al. (1994), Hutchins and Somers (1992), and in the collection edited by Newton (1992). For rule-based systems see the articles contained in King (1987), Nirenburg (1987) and Slocum (1988). The history of MT is covered by Hutchins (1986, 1995).

References

Arnold, D., L. Balkan, R. Lee Humphreys, S. Meijer, and L. Sadler. 1994. *Machine Translation: An Introductory Guide*. Manchester: NCC Blackwell.

Hutchins, W. J. 1986. *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood.

——and H. L. Somers. 1992. *An Introduction to Machine Translation*. London: Academic Press.

——1995. 'Machine translation: a brief history'. In E. F. K. Koerner and R. E. Asher (eds.), *Concise History of the Language Sciences: From the Sumerians to the Cognitivists*. Oxford: Pergamon Press, 431–45.

King, M. (ed). 1987. *Machine Translation Today: The State of the Art.* Proceedings of the Third Lugano Tutorial, Lugano, 2–7 Apr. 1984. Edinburgh: Edinburgh University Press.

Newton, J. (ed). 1992. *Computers in Translation: A Practical Appraisal*. London: Routledge.

Nirenburg, S. (ed). 1987. *Machine Translation: Theoretical and Methodological Issues*. Cambridge: Cambridge University Press.

Slocum, J. (ed). 1988. *Machine Translation Systems*. Cambridge: Cambridge University Press.