[*Terminologie et Traduction*, no.2, 1985]

SYSTRAN MACHINE TRANSLATION PROCESSING AS AN EXAMPLE OF NATURAL LANGUAGE UNDERSTANDING*

Ian M. Pigott
Commission of the European Communities

Introduction

At this congress on terminology and knowledge engineering, I thought it would be fitting to present the mechanics of the Systran approach to language understanding on the basis of a concrete example of the actual machine translation process.

For those of you who require additional information on the history of the system and how it has been developed at the EC Commission for European language combinations, I would refer you to our current status report which is available here today.

Suffice it to say at this stage that the Commission has been involved in Systran developments since 1976 and that we now have nine operational language pairs, six based on English as a source language and three on French. We hope to embark on German as a source language in the very near future.

Elsewhere in the world, Systran has been applied to a wide variety of other language combinations including source and/or target modules for Russian, Japanese and Arabic. Recently, an agreement on joint development was concluded between the Commission and Gachot S.A. of Soisy near Paris who own three major Systran development groups (Systran Institut, World Translation Center and Latsec). In addition, a Japanese company, Iona, has a subsidiary called Systran Corporation which is developing language pairs involving Japanese.

Technical aspects

The Systran software, which contains many hundreds of thousands of lines of IBM 360 Assembler programming as well as dictionaries running to well over 100,000 entries per source language, runs on IBM mainframes, generally under the MVS operating system. Access from local or remote workstations (e.g. personal computers or word processors) is achieved via a variety of batch or interactive telecommunications facilities. On an average size mainframe, the system will translate some 500,000 words per CPU hour and, in our environment, typical total elapse time for a 10-page document is around two minutes. ;

Direct computer capacity costs average about US $2 per page but the total cost to outside users may well be rather higher as a number of support services are usually involved.

* This paper was originally presented at the conference on Terminology and Knowledge Engineering, Trier, September 1987.

The translation process

There are ten main steps in the Systran machine translation process. I shall be explaining each of these in more detail later, but the following slide will give you a general overview:

SLIDE 1

Page format recognition

Source language analysis including
- Grammatical homograph resolution
- Clause boundary establishment
- Sequential parsing passes
- Deep structure representation

Bilingual transfer including
- Lexical routines
- Contextual dictionary rules

Target language synthesis including
- Creation of target morphology
- Reordering of words at target level

Re-establishment of page formatting

In order to illustrate some of the pertinent features of each of these stages, I have chosen the following sentence from a recent report on Systran. (By sheer coincidence, it happens to present some of the system's most interesting features.)

The sentence runs as follows:

SLIDE 2

This statement bears witness to the fact that Systran has already met the five basic criteria (speed, economy, accuracy, accessibility and flexibility) for a wide variety of users, ranging from large organizations, like the U.S. Air Force, NASA, the Commission of the European Communities, the Xerox Corporation and General Motors of Canada, who undertook much of the initial development, to many new users who today access Systran via telecommunications networks such as Minitel in France or the COTEL facilities developed as part of the Esprit programme.

As you can see, it is a fairly long sentence and contains quite a variety of linguistic problems requiring treatment at various levels.

Page format recognition

Generally speaking, documents are formatted for display and printing purposes. Very often the page presentation will have a direct bearing on semantic relationships between the various units of text. Headings may be centered, an introductory clause may introduce a number of subparagraphs or tables may be used for presenting indexes or statistics.

While Systran relies to some extent on punctuation for text unit analysis, punctuation alone is not enough. The full stop (or period) is, for example, often used after abbreviations. Many "sentences" are recognizable only by the fact that there is line spacing (without a full stop) to highlight the text unit.

The first step in Systran processing is to analyse the page representation as a means of dividing the text up into so-called translation units which often, but not always, coincide with actual sentences.

My example today actually appeared as part of the following page of text:

SLIDE 3

SYSTRAN: A MACHINE TRANSLATION SYSTEM TO MEET USER NEEDS

It is fitting that we should open this session on commercial machine translation systems with a presentation on Systran, a pioneer among MT systems, whose users and developers are best characterized by their shared persistent belief in the proposition that machine translation is feasible - and not only feasible, but practical and useful. It is very encouraging to look around me today and see how many of you have indicated by your presence at this conference that you share this belief. Because Systran developers have always had as a goal the production of useful translations - and because Systran has been a production system since 1969 - I would like to focus my talk on the five qualities that make a system useful: speed, economy, accuracy, accessibility and flexibility.

As a brief introduction to Systran, I would like to quote from Prof. Juan Sager's closing remarks at the World Systran Conference held in February, 1986, in which he stated that Systran "is used more widely and by a greater range of users for a larger diversity of purposes than any other system currently in use." This statement bears witness to the fact that Systran has already met the five basic criteria (speed, economy, accuracy, accessibility and flexibility) for a wide variety of users, ranging from large organizations, like the U.S. Air Force, NASA, the Commission of the European Communities (CEC), the Xerox Corporation and General Motors of Canada, who undertook much of the initial development, to many new users who today access Systran via telecommunications networks such as Minitel in France or the COTEL facilities developed as part of the Esprit programme.

Systran was developed by LATSEC, Inc. and World Translation Center, Inc. in La Jolla, California; today there are additional development groups in Luxembourg, Paris and Tokyo. Systran now offers 15 operational language pairs. These include English into eight languages: French, German, - Italian, Spanish, Portuguese, Russian, Japanese and Dutch: French into English, German and Dutch; and Russian, German, Japanese and Spanish into English. English-Arabic is under development, while pilot systems exist for 6 other language pairs: German into French, Spanish and Italian: and Chinese, Portuguese and Italian into English. The subject fields covered are already too numerous to list here, while document types range from abstracts, technical reports, journal articles, and service manuals to minutes of meetings and newspaper articles, with the range of linguistic styles increasing dramatically as many new users gain access to Systran.

As you can see, the page presentation here is relatively straightforward. There is one centered heading followed by a number of paragraphs. The beginning of our own sentence is clearly indicated by the period-unquote (.") sequence and the subsequent upper-case T (This) and the end comes not only with a period but with a line change. Note however that within the sentence itself the sequence "U.S. Air Force" could have been misinterpreted as a sentence break. The page recognition peripherals contained sufficient information to analyse U.S. as an abbreviation and so no break was made.

Global idiomatic replaces

The next stage in the process is based on a special dictionary lookup which reduces some of the closely related word strings to one-word equivalents with clearly defined part-of-speech values.

SLIDE 4:

This statement bears.witness to the
fact that Systran has already met the
five basic criteria (speed , economy ,
accuracy , accessibility and flexibility)
for a wide variety of users , ranging
from large organizations , like the U.S.
Air Force , NASA , the Commission of the
European Communities , the Xerox
Corporation and General Motors of Canada,
who undertook much of the initial
development , to many new users who today
access Systran via telecommunications
networks such.as Minitel in France or
the COTEL facilities developed as.part.of
the Esprit programme .

You will see that "bears witness" has now been reduced to a one-word equivalent of a verb while "such as" and "as part of" can now be considered as prepositions. This helps with later processing.

Main dictionary lookup and homograph resolution

The text is now ready for true linguistic parsing. Each word in the sentence is looked up in the English source dictionary and receives all the basic morphological, syntactic and semantic information coded in the corresponding dictionary entry.

One type of information required in the very first stage of analysis is that relating to grammatical homograph possibilities. In all languages, many word forms can function as more than one part of speech. In English, with its lack of morphological inflections, homograph resolution is one of the most difficult levels of natural language analysis. In our sentence, the highlighted words are all grammatical homographs of one kind or another.

SLIDE 5:

This statement bears.witness to the
fact that Systran has already met the
five basic criteria (speed , economy ,
accuracy , accessibility and flexibility)
for a wide variety of users , ranging
from large organizations , like the U.S.
Air Force , NASA , the Commission of the
European Communities , the Xerox
Corporation and General Motors of Canada ,
who undertook much of the initial
development, to many new users who today
access Systran via telecommunications
networks such.as Minitel in France or
the COTEL facilities developed as.part.of
the Esprit programme .

As you can see, in this sentence, 35 out of 86 words (or 41% of the words in the sentence) present homograph problems. THIS can be a pronoun or a demonstrative adjective, BEARS can be a noun or a verb, WITNESS can be a noun, a verb or an infinitive, TO can be a preposition or an infinitive particle, and so on. In fact, even words like THE and AND are potential homographs (THE = article or adverb, AND = coordinate conjunction or subordinate conjunction) but these are treated as special cases by so-called lexical routines.

Some homograph possibilities have already been taken care of by the global replace mechanisms already mentioned (e.g. BEARS immediately followed by WITNESS has been reduced to the equivalent of a verb group). Others, such as AIR FORCE and EUROPEAN COMMUNITIES, are covered by another level of dictionary coding for noun expressions. The remainder are resolved by special homograph routines which conduct a variety of syntactic tests on the surrounding context. One type which is particularly difficult to solve is the past participle vs past tense resolution for most English verbs (MET and DEVELOPED in this sentence, and preposition vs verb (LIKE).

I am happy to report that all grammatical homographs were successfully resolved in our example.

Clause boundary analysis

The next stage of processing, now that the part-of-speech functions of each word have been established, is clause boundary analysis. The aim here is to establish the main clause of the sentence, any subordinate clauses and any groups such as infinitive clauses or bracketed information which require special treatment.

SLIDE 6: This statement bears.witness to the fact that Systran has already met the five basic criteria (speed , economy , accuracy , accessibility and flexibility) for a wide variety of users , ranging from large organizations , like the U.S. Air Force , NASA , the Commission of the European Communities , the Xerox Corporation and General Motors of Canada, who undertook much of the initial development , to many new users who today access Systran via telecommunications networks such.as Minitel in France or the COTEL facilities developed as.part.of the Esprit programme .

As you can see from the colour coding, a number of clauses exist. The main clause (in blue) introduces the sentence while one of the subordinate clauses in fact causes a break in another of the subordinate clauses. For the purposes of subsequent processing, structural relationships need to be established for each clause as well as between the various clauses making up the sentence. After clause boundary analysis, we are in fact faced with the following clauses or sub-sentence units:

SLIDE 7: This statement bears.witness to the fact

that Systran has already met the five basic criteria* for a wide variety of users, ranging from large organizations , like the U.S. Air Force , NASA , the Commission of the European Communities , the Xerox Corporation and General Motors of Canada ,** to many new users***

*(speed , economy , accuracy , accessibility and flexibility)

**who undertook much of the initial development ,

***who today access Systran via telecommunications networks such.as Minitel in France or the COTEL facilities developed as.part.of the Esprit programme .

The sentence has now been broken down into four separate clauses or sub-sentence groups, each of which can be analysed in its own right.

Sequential parsing passes

The next stage in the analysis process consists of establishing grammatical relationships between the various words in each clause. Initially, only the closest and most direct relationships are established. For example in our main clause :

SLIDE 5: This statement bears.witness to the fact

STATEMENT is established as the subject of BEARS.WITNESS, TO is analysed as a preposition governing FACT. Relationships are also established between the demonstrative adjective THIS and STATEMENT, the noun it qualifies, and between the definite article THE and the noun FACT.

As the structures here are relatively simple, no major problems exist. However, if we take the first subordinate clause:

SLIDE 9: that Systran has already met the five basic criteria* for a wide variety of users, ranging from large organizations , like the U.S. Air Force , NASA , the Commission of the European Communities , the Xerox Corporation and General Motors of Canada ,** to many new users***

the establishment of relationships is somewhat more complex.

For example, HAS is not a verb in its own right but is an auxiliary modifying MET. In the structure THE FIVE BASIC CRITERIA, the noun is related to three qualifiers: the article (THE), a number (FIVE) and an adjective (BASIC). Pointers are first established for the closest relationship (BASIC and CRITERIA) and later between the other qualifiers and the noun.

Another major problem to be solved is that of enumeration. In this clause, the terms U.S. AIR FORCE, NASA, THE COMMISSION OF THE EUROPEAN COMMUNITIES, THE XEROX CORPORATION and GENERAL MOTORS OF CANADA are all in enumeration. The head words of each term (FORCE, NASA, COMMISSION, CORPORATION and MOTORS) must be clearly established and corresponding pointers are set between them. It is in fact no easy matter to decide that the term COMMISSION OF THE EUROPEAN COMMUNITIES is a unit in its own right and that the semantic affinity of COMMISSION does not extend further down the clause to XEROX CORPORATION, etc. To the human reader, it is quite obvious that there is no question of a COMMISSION OF THE XEROX CORPORATION, yet syntactically this is possible. The analysis here depends first and foremost on the semantic coding of the various head nouns which all carry a dictionary code notating the fact that they are organizations. In the case of GENERAL MOTORS, where MOTORS would normally be interpreted as a device, a special contextual dictionary rule based on the capitalization and the plural noun comes into play.

There are in fact four sequential stages of structural parsing which finally lead to the establishment of various levels of pointers between all the component words of the sentence. At a general level, the entire predicate of the sentence contains pointers indicating the main verb and its subject. The conjunctions and relative pronouns (THAT and two occurrences of WHO) are clearly related to their antecedents and, last but not least, a deep structural representation of the sentence is obtained.

In this way, it is possible to identify the deep subject (agent) of passive constructions and prepare for transfer into active forms in the target language. Such transformations are however not required in today's example.

I do not have time today to go into all the complexities of the numerous syntactic and semantic codes in the dictionary which are used during the analysis process. Let me just say that apart from standard morphological and grammatical information on part of speech, gender, number, case, person and tense, a large number of syntactic and semantic codes are used to define the behaviour of words in context. The syntactic codes serve to indicate potential dependencies such as "can govern an infinitive," "may introduce a noun clause," "always transitive," etc., while semantic codes provide markers giving information about the meaning class of a word. For example, a verb might be coded as normally requiring a human subject and a concrete object and being associated with motion. Nouns fall into categories such as concrete or abstract at an elementary level and, at a deeper level, carry markers such as DEVICE, PROFESSION, PROPERTY, etc.

Attached to each word is an extensive processing area (160 eight-bit bytes) available for storing and interpreting this information and establishing dependencies between each word and its contextual associates.

Thus, by the end of analysis, which is completely monolingual, a very rich representation of sentence structure is available for the subsequent stages in the process, transfer and synthesis.

Bilingual transfer

The main role of transfer is to deal with those areas of translation processing which go beyond the comparatively regular default-type processing in the target generation programs.

For example, routines to handle structures involving dates or the specific requirements of proper nouns such as place names are handled at this level. Fairly complicated routines also exist for handling the translation of terms or structures such as THERE IS, AS, EXPECT, etc. where the choice of target meaning and syntax depends heavily on context.

In our sentence, words such as THAT, WHO and IN are supported by lexical routines.

Another extremely important component at the transfer stage is the assignment of special meanings in context. A good example here is the translation of the word MET. The basic French translation of MEET is RENCONTRER. But in our sentence this meaning would be incorrect. In French, there is a close affinity between the verb MEET and its object CRITERIA which leads to a specific translation.

Insertion of the correct translation is facilitated by a dictionary rule which specifies - on the basis of the results of analysis - that when the direct object of MEET is CRITERIA, the French translation is REPONDRE followed by the preposition A.

Similarly, WIDE when qualifying VARIETY is not to be translated LARGE but GRAND. ORGANIZATIONS (in the plural) is more likely to mean ORGANISMES than ORGANISATIONS. U.S. in an adnominal relationship requires an adjectival translation (AMERICAIN) rather than the literal noun phrase DES ETATS-UNIS. All these cases are handled by contextual rules which do much to enhance the level of raw translation quality.

Target language synthesis

As an indication of what type of processing is handled in synthesis, let us look at what the word-for-word translation of our sentence would have been without any target morphology or reordering of words:

SLIDE 10:

Ce déclaration témoigner de le fait que Systran déjà répondre à le cinq de base critère (vitesse, économie, précision, accessibilité et flexibilité) pour un grand variété de utilisateur, se étendre de grand organisme, comme le américain air armée, NASA, le Commission de le européen Communauté, le Xerox corporation et General Motor de le Canada, qui entreprendre un grand partie de le initial développement, à beaucoup nouveau utilisateur qui aujourd'hui interroger Systran par intermédiaire de télécommunication réseau tel que Minitel en France ou le COTEL équipement développer en tant que élément de le Esprit programme.

The target synthesis programs create the correct inflections, add necessary articles, undertake reordering of words and handle any other syntactic requirements of the target language (elisions, infinitive particles, prepositions, etc.).

The final result is then as follows:

SLIDE 11:

Cette déclaration témoigne du fait que Systran a déjà répondu aux cinq critères de base (la vitesse, l'économie, la précision, l'accessibilité et la flexibilité) pour une grande variété d'utilisateurs, s'étendant des grands organismes, comme l'armée de l'air américaine, la NASA, la Commission des Communautés européennes, la corporation Xerox et la General Motors du Canada, qui ont entrepris une grande partie du développement initial, à beaucoup de nouveaux utilisateurs qui interrogent aujourd'hui Systran par l'intermédiaire des réseaux de télécommunications tels que le Minitel en France ou les équipements COTEL développés en tant qu'élément du programme d'Esprit.

While the translation is not perfect, it is good. The French can be understood perfectly without any reference to the original English. Were post-editing to be undertaken, it would be for purposes of improving style rather than correcting grammar or terminology.

I would argue that in many environments, the level of translation obtained here would be fully acceptable.

It is only fair to point out, however, that not all sentences translated by Systran come up to this level of quality. But many do, as can be seen from the various samples available here today.

Conclusion

I hope I have been able to demonstrate today that Systran is indeed an example of a system which has sufficient intelligence for achieving natural language understanding in the specific area of machine translation.

I would not assert that the system has anything approaching a world knowledge component, but this is hardly necessary for the more limited task of translation.

Competent human translators are also able to achieve excellent results in fields of science and technology about which they have only a limited level of knowledge. They do not need to master Einstein's theory of relativity to be able to translate nuclear research reports. What they do need is a good basic command of the source and target languages as well as a "store" of the relevant terminology.

Systran's "understanding" of natural language is at the same level. In today's example sentence, it recognizes, for example, that the word FACT is likely to introduce a noun clause and that SPEED, ECONOMY, ACCURACY, ACCESSIBILITY and FLEXIBILITY are all nouns covering properties or qualities. Drawing on information of this type, Systran is able to construe mathematically the kind of transformations a human being would bring into play when confronted with the same translation problems.

With up to 150,000 dictionary entries and target language equivalents per language pair, it also has a very large store of general and technical terminology.

To this extent, Systran's understanding bears close similarity to the level of understanding of natural language applied in the human translation process.

And this is hardly surprising, as unlike many other machine translation systems, Systran development has been coordinated by a team of translators who have been able to apply their own experience of the translation process to writing complex programs and dictionary rules which are the basis of today's system.