HOW SYSTRAN™ WORKS

The description below assumes that the reader is already versed in the basics of machine translation. For further information about the concepts and underlying principles referred to here, see Arnold et al. (1994), Hutchins & Somers (1992), or Vasconcellos et al. (1993).

Overview

SYSTRAN™, well known for its venerable history of service to government and industry, has a flexible architecture that has enabled it to evolve in pace with changing technology and emerging insights in the field of computational linguistics. Without losing the benefit of the hundreds of person-years invested since 1968 in the development of dictionaries and linguistic rules for its impressive range of source and target languages, SYSTRAN has been able to transition successfully into a mature *transfer-type machine translation system.*

Multitarget/multisource approach. Beginning with the English-French system in 1974, SYSTRAN has been designed to be multitarget. Multiple target languages can be attached to a single analysis module because the analysis is devoted exclusively to processing information about the source-language text; in other words, no information about target languages is handled in the analysis phase. In 1987 the modularity, consistency, and economy of SYSTRAN's romance-language analysis modules were further enhanced by combining many of the shared functions in a single multisource analysis. Development of a second set of shared analysis functions has begun for the Altaic languages, Japanese, and Korean.

The fact that SYSTRAN is a multitarget system is important to the modularity and maintainability of its many language pairs. It means that, for a given source language, development work on dictionaries and linguistic rules need only be done once and it will apply to all the target-language synthesis modules via a language-pair transfer module.

Once SYSTRAN completes its analysis of the source text, a symbolically represented output is passed on to the transfer module, which then applies a series of rules that set the stage to produce translations in the various target languages. The transfer component is characteristic of the architecture of a *transfer-type MT system.*

The final stage in the automatic translation process is the synthesis. This is the module that actually produces a text in the target language(s). In theory, the number of target languages is unlimited.

The three components—analysis, transfer, and synthesis—are described in greater detail below. Altogether, SYSTRAN has developed 10 linkable source-language modules that make for a total of 27 operational language pairs.

Knowledge sources. The basic job of any machine translation system is to store and make use of knowledge. SYSTRAN has two major knowledge stores: its electronic dictionaries, and the linguistic rules that interact with them. The dictionaries contain information about the behavior of each specific word, while the linguistic rules refer to the syntax of a language or sublanguage and to semantic relationships between concepts. The dictionaries contain masses of "bottom-up" rules about the particular requirements and preferences associated with specific words, while the linguistic rules work in the "top-down" direction to establish syntactic and semantic relationships. The vast amount of linguistic knowledge that is stored in the dictionaries and rule bases has kept SYSTRAN at the forefront of the MT industry for more than a quarter of a century.

Language-independent format. One of the reasons why SYSTRAN's approach can be so readily expanded, and why new language pairs can be developed quickly and reliably, is that the basic features are represented in the same manner in all of SYSTRAN's many language modules: its architecture, dictionary coding, symbolic

representation system, and "recipe" for analysis, transfer, and synthesis are all language-independent. It is this consistency that gives power and efficiency to SYSTRAN's mechanism for multilingual data retrieval.

Robustness. The SYSTRAN staff has always been committed to developing robust systems capable of handling large volumes of general text that was not prepared with machine translation in mind. When SYSTRAN is used for this kind of input, it may happen that some of the source text is ill-formed, nonstandard, or even corrupted in some way. In the event that the parse fails, SYSTRAN's localized bottom-up rules still allow it to produce target output. Even the elements and phrases of an incomplete sentence can be analyzed and successfully synthesized.

When a word in the source text is not found in the stem dictionary, the first step is to search for alternative spellings, including variants both with and without orthographic accents. If a match is still not found, the system then attempts to determine the word's function based on its morphology and on the immediate context.

SYSTRAN also has an error-flagging program that can be activated at the end of the analysis module. When a parse fails, a flag sends a signal to the transfer module to ensure that paths will not be followed which will compound the problem further.

Document type. While SYSTRAN's priority is to be able to handle general-purpose texts, different text types have variations in lexical usage and grammatical conventions which mean that one set of linguistic rules cannot fit all. The type of document may be specified by the user at run-time either for an entire translation job or for portions thereof. The following document types are available: abstract, business correspondence, newspaper, patent, parts list, instruction manual, minutes proceedings, prose, and conversational/ colloquial text.

This option is implemented through switches at various points in the analysis which selectively adjust the rules to the characteristics of the different text types. In addition, certain stylistic choices are made in the synthesis stage.

Electronic Dictionaries

SYSTRAN's large and heavily encoded dictionaries are fundamental to its translation capability. For each source language there are two dictionaries: the *stem dictionary*, and the *expression dictionary*. Most of the source-language dictionaries are multitarget. As mentioned earlier, the 10 source-language modules may be combined to produce 27 language pairs. Altogether, the dictionaries contain a total of more than 2.3 million fully encoded words and expressions.

The dictionaries are not separated into domain-specific databases. Domain-related differences are handled by means of identifying codes in the source-language dictionaries, which work in tandem with alternative meanings in the target language corresponding to different domains. The system selects the domain-specific meanings according to the particular *topical glossaries* requested by the user at the time the translation is run. From one to four topical glossary selections, stacked in order of preference, may be specified at run-time.

Stem dictionaries. The stem dictionary contains the base or uninflected forms of single words. Each word is accompanied by extensive encoded information about its morphology, syntactic behavior, the possible functions it may perform if it is homographic, semantic roles, and semantic attributes and relationships to other concepts based on a 500-category semantic taxonomy.

Target-language meanings are provided for as many languages as the developer wishes to include. The codes assigned to them refer to part of speech, morphology, syntactic behavior, and prepositional government. For each polysemous word (i.e., homograph within the same part of speech), multiple meanings may be assigned for different domains and for different uses of the word (for example, animate/inanimate use in the case of nouns, transitive/intransitive or reflexive/nonreflexive use in the case of verbs).

Expression dictionaries. The expression dictionary may include several types of entries, which are listed below in order of increasing complexity.

The *idiom replace* allows frozen idiomatic expressions and multiword prepositions or adverbial phrases to be fused into a single *pseudo stem*, which is then entered in the stem dictionary as a single word. It is parsed as a single token.

The *collocation* assigns a single meaning to a phrase, the elements of which are parsed and inflectible. It is useful for conventionalized technical noun phrases, e.g. "lug nut."

The *conditional expression* is enlisted when meanings or other target-language information should be invoked under certain conditions only. The conditions for meaning assignment may utilize any of the syntactic criteria, including syntactic features, or semantic attributes or relationships that have been defined within SYSTRAN, and the rules can be quite elaborate. They are called in at the transfer stage to select the target meaning and perform other transfer-phase operations such as syntactic reordering and adjustments in prepositions, determiners, tense, etc., to reflect the requirements of the target language.

The *parsing expression* applies word-specific rules in the course of the parsing process. It is especially useful for early disambiguation of polysemous words or those that have multiple patterns of syntactic usage. Any information from the source-language dictionary may be modified, including semantic attributes. Rules and information may also be added in this way. A parsing expression may be invoked at any point in the analysis.

The *homographic expression* disambiguates and assigns the correct part of speech to a single word.

Further details about the SYSTRAN dictionary structure and expression types are given by Wheeler (1983, 1987).

The Sequence of Events

Control software: processing of input. SYSTRAN's input module includes filters for a wide variety of word processing and desktop publishing formats. Format codes are separated from the text before it is sent for translation and then held in reserve for reattachment later.

The input text is processed one sentence at a time. The dictionary lookup routine performs morphological analysis and identifies capitalization, punctuation, and hyphenation. After the lookup is completed, the input module also assigns a part of speech to any word not found in the dictionary based on its morphology and immediate context.

Analysis component. The analysis module goes systematically through the sentence, gradually identifying the correct function and meaning of every word, phrase, and clause by means of a series of passes. Each of these passes makes decisions or inferences about a particular type of syntactic or semantic phenomenon—for example, resolution of ambiguities and basic syntactic relationships, prepositional government, semantic relationships, clause boundaries and clause types, coordinate constructions, etc. (for an in-depth description, see Wheeler 1987). Each pass adds new data to the information being accumulated about the sentence. As the knowledge is captured, it is saved in the *analysis area* in the form of a symbolic representation. Again, it should be noted that throughout the entire analysis phase the knowledge accumulated refers to the source language only.

Over time SYSTRAN has drawn on different linguistic theories in its process of scaling up to increasing levels of complexity. However, the symbolic representation always expresses the same phenomena identically across languages.

The basic parse of a sentence forms an extensive tree of relationships. These up-down links indicate government, dependency, and modification. During the analysis, moves between constituents are made via these links. If a graphic tree were to be generated from the symbolic representation, it would resemble a dependency tree. The nodes would carry extensive information about the function and type of relationship between each constituent in the clause and also between the matrix and subordinate clauses within a given sentence. For each clause the parse identifies the head noun of the subject noun phrase and also the main predicate.

One of the tasks of the analysis component is

to capture and save certain information about the subject and predicate of the current sentence for reference later on.

In addition to the syntactic analysis, the following semantic roles are also identified: predicate-agent, predicate-patient, and head-modifier. These roles are used to supplement syntactic information in linking constituents.

At the same time, SYSTRAN's 500-category semantic taxonomy provides information about the characteristics and relationships of things, actions, states, and qualities which can be useful in making decisions about the behavior of words or the objects they represent. The individual semantic categories, represented by *tags*, may be assigned to words or phrases either in the stem dictionary or through a general linguistic rule.

The taxonomy is organized around six hierarchical trees. In general, lower nodes of the tree inherit all the properties of the nodes above them, although inheritance may be blocked when it is desirable to do so.

Transfer component. One of the main functions of the transfer module is to handle grammatical dissimilarities between different languages. As part of this task, SYSTRAN may alter or rebuild clause and phrase structures in order to meet the syntactic requirements of the target language.

The second main function of the transfer module is to select a target-language meaning for a source-language word that calls for such a decision. Extensive use is made of linguistic rules and dictionary expressions. The many syntactic and semantic relationships that have been established up to this point allow for a wide range of tests to be applied to the constituents. Another important aid for disambiguation is the abundance of semantic information that has been attached to the words in the surrounding context.

Also during the transfer phase additional lexical rules may be applied to classes of words in order to adjust tense, aspect, number, voice, or any other feature. Such adjustments play a key role in ensuring that the target output is grammatical and naturally phrased.

Synthesis component. Because SYSTRAN uses a transfer approach, the final synthesized output tends to be fairly faithful to the syntax of the source language.

The synthesis module assigns inflections for case, number, tense, and aspect based on the information derived from the analysis together with the syntactic requirements of the target language. For the latter purpose, a large number of rules and tables are provided which apply specifically to the target language.

This module can also insert or delete determiners, including definite and indefinite articles, as well as other particles.

Control Software: Processing of Output. At the end of this sequence, the control routines retrieve the format codes that were separated out and saved at the beginning, and they now reattach them to the output. Finally, they print out the target text sentence-by-sentence.

REFERENCES

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

Hutchins, W. John, & Somers, Harold L. (1992) *An Introduction to Machine Translation.* London, New York, etc.: Academic Press.

Vasconcellos, M.; Hovy, E.; Scott, B.E., Miller, L.C. (1993) "Machine Translation: State of the Art." *Byte*, January, pp. 153-186.

Wheeler, Peter J. (1983) "The Errant Avocado." *Newsletter of the British Computer Society, Natural Language Translations Specialist Group*, 13.

Wheeler, Peter J. (1987) "SYSTRAN." In King, Margaret, ed., *Machine Translation Today: The State of the Art.* Proceedings of the Third Lugano Tutorial (Lugano, 2-7 April 1984). Edinburgh: Edinburgh University Press. Information Technology Series 2. pp. 192-208.