TRANSTAR —

A COMMERCIAL ENGLISH-CHINESE MT SYSTEM

Dong Zhen Dong

Language Engineering Labs
China Software Technique Corp.

TRANSTAR is widely acknowledged as the most successful and powerful MT system in China. It has won favourable reactions from the demostrations in Singapore (1986), Hong Kong (1987) and Hanover (1989). The system has been sold to over 50 users both at home and abroad since its release in September 1988. The development of the system, from basic design to commercialization, took more than 11 years.

1. Current Specifications

The current specifications of the system are as follows:

— Dictionaries

.Basic dictionary: 45,000, including words,phrases, idioms and proverbs etc.
.Technical-term dictionaries: 200,000 technical terms, covering a wide range of subject fields, such as economics, computational technology, telecommunications, car manufacture, acoustics etc.

— Hardware

The system can run on the following machines: IBM AT and its compatibles, VAX series, Wang PC, IBM S/20, Olivitti PC, Universe 68000, etc, with COBOL as its programming language. The development of a new version on UNIX with C language is being under way.

— Translation Speed

The translation speed varies greatly with different types of machines, approximately ranging from 1,000 words/hour to 3,000 words/hour.

— Translation Accuracy

Generally, the translation accuracy, when no pre-editing and post-editing are applied, is "OK", as the sinologists claimed at Hanover Show in 1989. The accuracy depends much on the subject fields or styles of the texts to be translated. For example, the translation is terribly poor for news reports.

— Facilities for End-users

. Input or output document management tools: to be used to help users to do pre-editing (though minimal), and post-editing in two-column display.
. Dictionary maintenance tool: to be used for users to amend the dictionaries in the following ways: <1> to add new words or terms by giving a template, rather than manual coding for detailed dictionary information; <2> to update Chinese equivalents to the English enteries, if they are thought to be unfit to the user's taste.

2.Overview of TRANSTAR

TRANSTAR is a transfer-based, batch-operating unidirectional English-Chinese MT system. The configuration and the operation process of the system are outlined in Fig. 1.

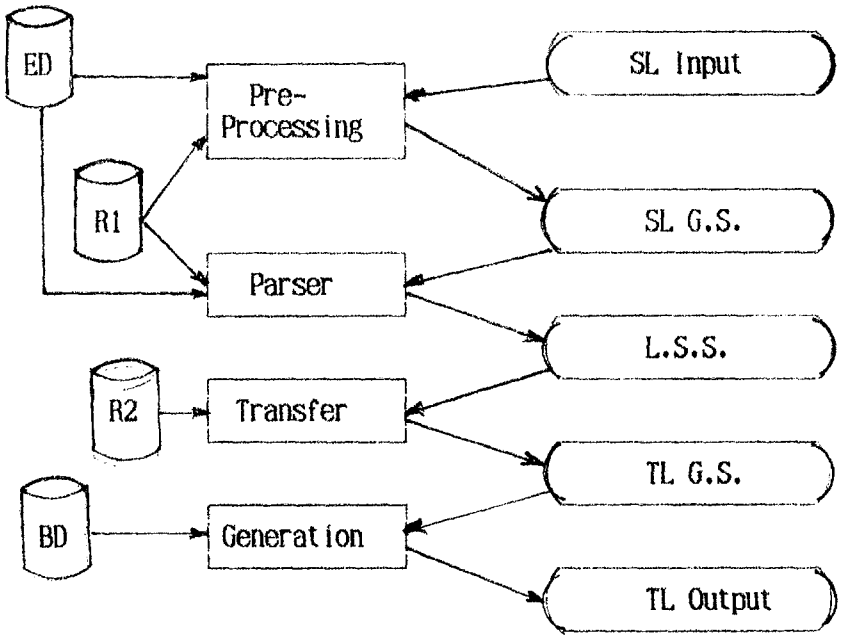
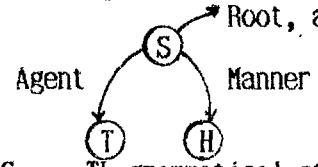


Fig.1

ED — English monolingnal dictionary
BD — English-Chinese bilingual dictionary
R1 — Rule-type 1 used for analysis
R2 — Rule-type 2 used for transfer
SL Input — as "They studied hard."
SL G.S. — SL grammatical structure, as "They study hard."
L.S.S. — Logical semantic structure, as



TL G.S. — TL grammatical structure, as Agent < Manner < Action
TL Output — as "他们努力学习。"

3. Dictionary Configuration and Information

In TRANSTAR, the basic dictionary or each of the technical-term dictionaries consists of two sub-dictionaries: an English monolingual one and an English-Chinese bilingual one. Any technical-term dictionaries should be used in combination with the basic dictionary. When the system runs, it should use two operating dictionaries. The dictionary configuration in TRANSTAR is shown in Fig. 2.

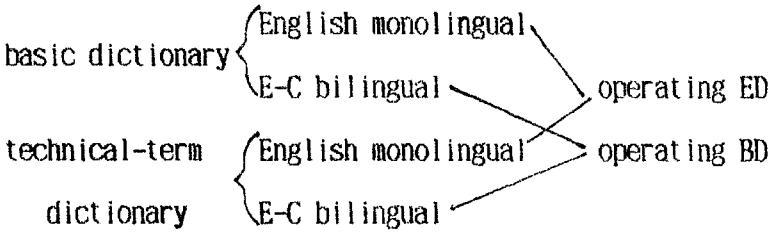


Fig.2

The internal structures and information types for both the basic dictionary and technical-term dictionary are just the same. The information given to each entry is allocated in different information zones: morphological, syntactic, semantic and transfer zones. Each Entry in ED's may have 26 information items. In the system's dictionary compiling manual, the information specified totals about 630 items, including 390 items relating to morphology and syntax, and 240 items to semantics.

4. Rule-base

All the rules of TRANSTAR are stored in two rulebases. One of them is a syntax-driven rulebase (SDRB), and the other, lexicon driven rulebase (LDRB). The former deals with general problems in the process, while the latter takes care of specific linguistic phenomena, such as idiomatic usage of words, verbal phrases, and ambiguity which can only be solved by means of semantic analysis, etc. The rules in SDRB currently total about 3,200, and in LDRB, about 2,600. Moreover, TRANSTAR's rules can be divided into two types in terms of their use in the operation of the system, Type 1 is used for analysis, and Type 2 for transfer. The rules in SDRB are organized in 10 main modules with 86 submodules included (See Table 1).

Functions	Submodule Numbers
Unknown-word treatment	3
Homograph disambiguation	21
Predicate or verbal determination	5
Clause scope marking	1
Phrase construction	7
Co-ordinate construction	4
Logical-semantic relation	9
Preposition phrase disambiguation	28
Transfer for simple sentences	7
Assembly of clauses	1

Table 1

All the rules of TRANSTAR are production-rules, written in a problem-oriented language — SCOMT. By using SCOMT, the separation of linguistic data from algorithm is achieved. Each rules is identically composed of 4 parts: <1> index which indicates which sub-module it belongs to and when it should be called; <2> condition, either simple or complex; <3> action which may include a set of actions conducted actually by subroutines; <4> next-step instruction which indicates what is to be done after the present rule is executed.

5. Parser and Logical Semantics

TRANSTAR uses a HPSG-like formalism, Constituent Functional relation Grammar (CFRG). The parser is characteristic of CFRG, an intermediate list structure will be generated, which represents two types of structure: syntactic structure and logical semantic structure. The nodes in both the structures are connected in a two-way mode. Thus the mother node can be visited through her daughters, and vice versa. The information items for each node will increase from 26 initial ones taken from thedictionary to 60 to the maximum, including static features and dynamic features. The logical semantic structure, similar to a case-structure serves as the basis of transfer. Logical semantics is represented by two types of semantic feature, one of which is the semantic relation between the concepts such as agent, patient, manner, duration, etc. The other is the attributes on the concepts,