System Demonstration

SYSTRAN Chinese-English MT, ECI Chinese OCR - CHIOCR

Laurie Gerber

SYSTRAN Software Inc.
7855 Fay Avenue, Suite 300
P.O. Box 907
La Jolla, CA 92037
tel: 619459-6700
fax: 619-459-8487

Introduction:

Development of the SYSTRAN Chinese-English system began in August 1994 at SYSTRAN Software, Inc., in La Jolla, California. This demonstration is based on an integration test done at SYSTRAN Software, Inc. as a beta test site for CHIOCR. Additional background and comments on the integration appear at the end of this paper.

System Builders and Contacts:

SYSTRAN Software, Inc.- SYSTRAN Chinese-English MT

Project Manager	Laurie Gerber - laurie@systranmt.com
Marketing Manager	Reba Rosenbluth - reba@systranmt.com
Phone	(619) 459-6700
Fax	(619) 459-8487

System Category:

SYSTRAN Chinese-English MT is under development for clients in the U.S. government, and until the pilot release, is available only to development sponsors. The pilot system and other subsequent releases will be made commercially available.

Alpha MT system release:	December 1995
Commercial pilot release:	End of January 1997
All-encoding-schemes release:	End of May 1997

Syste m Characteristics:

Domains covered: Technical and general text

Input format accepted: Traditional (Big5) and simplified (GB) Chinese, plain text format.

Output Quality:		
	Alpha release	~50% accuracy* for technical or general text
	Pilot release	Over 70% accuracy* for technical or general text

*In the experience of SYSTRAN users, achievement of 70% accuracy, as defined in SYSTRAN's internal quality analysis method, is the point at which an MT system become a useful productivity tool, i.e., when post-editing the output is faster than translating from scratch.

Dictionaries:

Chinese-English dictionary:		
	Alpha release	64,376 words and compound word
	Pilot release	~120,000 words and compound words

Hardware Platforms and Requirements:

(Items in parenthesis are only necessary for OCR integration)

Hardware: IBM compatible PC - 386 or faster processor
Super VGA+ monitor (for character display)
(HP ScanJet or other TWAIN compliant scanner)

Disk space: 20 - 200 MB free:
(15-18 MB Chinese Windows)
(10 MB Chinese OCR)
20 MB Chinese-English MT
(150 MB for .tif (scanned graphic image) files)

Memory: 16-20 MB RAM (20 MB if all applications (OCR & MT) may be kept running at the same time.)

Operating system:

SYSTRAN Chinese-English:	English or Chinese Windows (3.x, 95, NT), Unix
CHIOCR:	Chinese Windows 3.x

The Marriage of OCR and MT:

Much of SYSTRAN'S foreign language->English development has originally been done for clients who need to collect and assimilate information. The Chinese-English system is no exception. Such clients are faced with the challenge of a huge volume of foreign language material they would like to access, a limited number of qualified linguists and translators, and long turnaround times for translations or abstracts. At the same time, the quality requirements for output are often rather modest - lower than what would be produced by most human translators. The user often just needs to filter the useful from the useless, or get a general idea of content. MT has long been seen as the perfect solution for this type of translation bottleneck.

However, a second bottleneck immediately appears - much of the foreign language material of interest is not available in electronic form. For many years, the U.S. Air Force has worked around this with a production-line approach to input and translation of large volumes of material. But that method still is very labor intensive and requires well-trained typists. It can't work for someone in the field who encounters foreign language papers and needs to know quickly if they have found something important.

As a solution to these obstacles, input via OCR is now receiving a lot of attention. In autumn 1995, CHIOCR was beta tested at SYSTRAN. In addition to tests performed by the Chinese linguistic staff, a test was run involving a non-Chinese reader. (One objective of automating the input process is to allow people who only know the target language to do it.) Given the highly visual interface for OCR output editing, this test worked quite well, and the non-Chinese user was able to correct over half of the errors marked "uncertain" by the OCR.