LANGUAGE INDUSTRIES SURVEY

Language industries survey

Undersøgelse af sprogindustrierne

Untersuchung über die Sprachindustrie

Encuesta sobre las industrias de la lengua

Enquête sur les industries de la langue

Inchiesta sulle industrie della lingua

Enquête over de taalindustrie

Inquérito sobre as indústrias da lingua

ευνα σχετικά με τις σιομηχανιες της γλώσσας

introduction

This is the last in a series of six Language Industry Survey (LIS) Bulletins to have appeared in Electric Word magazine. The six LIS Bulletins have focused on the demand side of language technology, examining how companies have implemented new technologies to solve problems in dealing with language and paperwork.

For the preceding five LIS Bulletins, we've talked to users about their applications of digital image processing, optical character recognition, text-to-speech, CD-ROM and terminology management within their businesses. In this, the final LIS Bulletin, we report on an impressive machine translation installation in Canada.

To reiterate a final time, the fundamental theme of the LIS Bulletin has been our preception that the daunting multilingual challenge posed by the European Single Market is a unique opportunity for Europeans to acquire valuable skills in developing technologies to meet this challenge. If European businesses rise to the occasion, we believe they will be supremely positioned to do business in the rest of the world.

Computerized translation, known commonly as Machine Translation (MT), is the most unequivocally complex and most exalted of the new language technologies. As such, it is a fitting finale to the LIS Bulletin.

Whether fully automatic, high quality translation (FAHQT) is ever achieved or not, the exercise of trying to teach computers to translate is giving us valuable insight into human language, computers and the interface between humans and computers.

MACHINE TRANSLATION

A LITTLE BACKGROUND ON MACHINE TRANSLATION (MT)

Since the late 1940s, interest in applying computers to the task of translating has simmered with varying levels of activity. During the 1950s, much excitement and research flourished in America, Europe, Russia and other parts of the world. After a lull in the 60s and early 70s, activities revived in the late 70s, albeit with more pragmatic expectations than before.

Interest in MT is so substantial because of the enormous documentation burdens faced by companies wanting to do business in foreign countries as well as companies saddled with domestic multilingual obligations. Interest is most notable in Japan, but also in North America, Europe, and other parts of the world.

Currently, there may be a dozen commercial MT systems on the market, with perhaps twice that number of significant research projects in the wings. The commercial offerings range from simple PC-based mechanical dictionaries to sophisticated mainframe-based systems of considerable power.

The decision to implement an MT system requires careful assessment company needs, as well as the nature of the documentation to be translated. Whatever the circumstances, MT systems require a substantial investment in money, time and manpower.

A CASE STUDY

problem

A government contractor in Canada faces the burden of translating 400,000 pages of technical documentation. By constitutional law, all documentation in Canada must be in both English and French.

Saint John Shipbuilding is a shipyard in Saint John, New Brunswick, in eastern Canada. It is a holding of the Irving conglomerate, which has diversified regional interests in oil, pulp and paper and shipping.

Saint John Shipbuilding is currently fulfilling a contract to build 12 frigates for the Canadian Patrol. Initial estimates four years ago indicated that the frigates would require 400,000 pages of documentation – in both English and French. To translate this by conventional means would require "horrendous" amounts of time and money. The yard estimated the average translation house could handle 10,000 pages a year. Vice President of Finance Ron Fournier posed the $64,000 question: "What can we do about this?"

Together with systems analyst Larry Rogers, Fournier set out to investigate the possibility of automating the translation of this documentation.

solution

After months of researching possible solutions, Fournier and Rogers settled on Logos, an MT system developed by Logos Corp., Dedham, Massachusetts, USA. They felt it was best suited for handling the target language, Canadian French.

Fournier and Rogers conceived the establishment of a translation and document processing center built around Logos to handle the voluminous bilingual documentation of Saint John Shipbuilding.

In August 1988, after an 18-month gestation period, this documentation centre, called Lexi-Tech, went into operation in the city of Moncton, a 50% English, 50% French city in northern New Brunswick. Currently, Lexi-Tech has a staff of 60, of whom half are translators and terminologists.

how and why

Cesar Pinto, Lexi-Tech's Director of Technical Services, describes their installation: "Basically, we're offering a total solution for the company's documentation needs. The documentation is received in a variety of forms, and we deliver camera-ready copy, English in the left column, French in the right.

"It would be great if all documentation arrived in digital form, but not all of the many subcontractors – the Taiwanese firm supplying deck planks, for example – use wordprocessors. The first stage, then, is what we call 'data capture.'

"Here, printed manuals and technical diagrams are scanned in. We have scanners from Kurzweil, Caere, and one from Anatech, which can scan documents up to 40 inches by 19 feet in size – these are ships remember, so some of the drawings are huge. This Anatech scanner is the only one in Canada. "The next stage we call pre-publication, or preprocessing. Here, all incoming text is gathered together and formatted in Interleaf, running on our cluster of 40 VAXs. This is strictly a technical stage; there's no pre-editing done.

"For the final stage, which we call MT, the text is uploaded to Logos, which runs on our IBM 9370 model 60 mainframe. Here, it is translated into French, then downloaded back to the VAXs. Interleaf automatically re-formats it in double columns, English and French. The translation is proofed by a translator, then the whole lot is printed."

results

So, how many of the 400,000 pages have been translated? Cesar Pinto: "We're about a third of the way through, and we're exceeding our expectations. Our first year, we expected to do 10,000 pages, yet managed to do 15,000. Already this year, we've done 40,000. Actually, the shipyard can't supply the documents fast enough. But we have other customers, too: Northern Telecom, New Brunswick Telephone, Bell, Digital, among others.

"Terminology building remains an important activity," emphasizes Pinto. "We've added about 150,000 terms and specific inferences to Logos' database. We have fulltime terminologists to do this – it's really a specialized profession. We compile company and branch-specific terminologies, but they're arranged hierarchically. If the system doesn't find a term on one level, it looks to the one above."

conclusion

Pinto says the system is a success. "Yes, it's serving its purpose well. We've had occasional problems with the physical limits of the Logos database, but those have been solved. We estimate it achieves an accuracy rate of 70 to 80 %.

"For a 10-to-20 page manual, it's not really worthwhile to run it through the system. In this case, we give it to a translator, and they consult the terminology database manually. But for books of 300 to 400 pages: we can just crank them out!" Based on Lexi-Tech's experience, at what point does Pinto think an MT rig like Logos justifies itself? "I estimate the threshold to be about 10,000 pages a year." He qualifies this by adding: "If you're doing 10 books, the first book will take twice as long. But the remaining books will be done in 1/10th the time."

The LIS Bulletin is a publication of the Language Industry Survey, an officially funded project of DG XIII of the Commission of the European Communities. The Language Industries Survey is a research and evangelical project intended to inventory and foster awareness of the language industries in Europe and around the world. If you are a researcher, developer or producer of hardware or software for the language industries, please write to us for a questionnaire. If you are a qualified respondent, you will receive a database of public information on a floppy disk about all the respondents (expected to number over a thousand), in addition to a free, one-year subscription to Electric Word. Send your request to:

LIS Questionnaire – EW #18
P.O. Box 5477
1007 AL Amsterdam
The Netherlands