Machine Translation at NAIC - A Perspective

DALE A. BOSTAD, USAF

The National Air Intelligence Center (NAIC), better known throughout much of its history as FTD (Foreign Technology Division) has a long and distinguished history as a translation organization and has gained some acclaim as perhaps the foremost operational user of machine translation (MT). NAIC recently celebrated more than 30 years as a user and developer of machine translation capabilities.

NAIC has long been associated in many peoples' minds as the user of SYSTRAN Russian MT during the Cold War. But the Cold War is over and the Department of Defense and the Intelligence Community have entered a new area. This is a time of dynamic changes, reduced budgets, reductions in personnel, reorganizations, and redefinitions of missions. MT and uses of MT are also being closely looked at. Despite the tight money, the prospects for MT look very good.

So with some 30 years of experience under our belts and the Cold War over, I think this is a good time to reflect on where we have been, where we are going, and what lessons we have learned on this long journey.

Translation Tradition Leads to MT

Translation, and in particular large volume translation has always been an integral component of NAIC and its predecessor organizations. As T-2 Intelligence in 1945-1946 over 1,500 tons of Nazi German aeronautical research and development literature for the 1933-1945 time period was cataloged, indexed, microfilmed, and eventually translated. The technical knowledge gained from these documents revolutionized American industry, especially in such fields as rocketry, magnetic tapes, night vision devices, liquid and solid fuels, and vacuum tube technology. As the World War II-related material exploitation and document translation programs closed and the Cold War grew hotter, technical intelligence efforts turned increasingly toward the emerging technological threat posed by the Russians.

Faced with an overwhelming translation burden and encouraged by the burgeoning success of computers, in 1955 FTD requested RADC in Rome, New York to develop an MT capability. The result of the effort was the installation of the IBM Mark II system in 1963. FTD continued its pursuit of newer and better things, culminating in the installation of the SYSTRAN Russian-English system in 1969. The success of the Russian system spurred development of German and French systems in the early 80's and expansion to new language pairs in the 90's. NAIC'S long tradition of translation and commitment to MT has never waned over the years.

From Rough to Raw

Obviously 30 years provides ample opportunity for experimentation and fine-tuning of procedures. As MT became more acceptable to users and as the systems got better, especially the Russian system, our procedures evolved to meet changing contingencies.

In the 1960's and the first years of the 1970's the Russian-English system was used to produce a very rough hard-copy translation. Post-editors edited the translation extensively - almost rewriting it - and then the edited version was sent to a recomposition pool where it was retyped. Alternatively, some translators used the rough machine translation as a "pony" when they dictated the translation on tape. The translator picked and selected those parts of the machine translation that were valuable - certain technical terms and phrases - and worked directly from the Russian text. The dictated translation was then transcribed. Machine translation was really only used as an aid to the translator. The end product was a very accurate finished product.

Because this method was slow and productivity was low (there were threats to disband translation operations) a radical step was taken and a very lightly edited, hard-copy MT translation process was instituted in 1974. Editing changes were written in ink on large fanfold computer printout. The degree of editing was at the discretion of the editor, but speed was the name of the game. The end product could only be called crude and the format primitive - a utilitarian product whose one saving grace was that it got the information quickly to the requester and allowed an immense backlog to be eliminated. This was FTD'S first effort at producing what we call "partially edited machine translation" (PET). Thankfully, this first period of producing PET'S lasted only two years when new computers allowed us to take a gigantic step forward.

From 1976 to the present a standard product, with some variations, has stabilized. This is on-line computer revision of machine translation using EDITSYS.

EDITSYS is a software module called at the end of Russian-English analysis that identifies certain potential problem areas in the output and brings these conditions to the attention of the post-editor. The post-editor must react to the highlighted condition and verify the accuracy of the machine translation version or make a revision. Post-editing is thus determined by a software program which automatically defines the minimum amount of post-editing, although the editor can do more. This type of software-driven post-editing has been our mainstay for nearly twenty years. The translation is then printed on high-quality paper on a laser printer and graphics and tables are merged in.

In 1987 NAIC developed a new MT application which we call "interactive machine translation". This system gives all users individual access to MT at their own terminals and has proved to be a defining moment in moving MT from a strict translation environment. The system is now available on some 1500 PCs within NAIC. This is raw machine translation for quickly gisting the significance of information. It is most frequently used for rapid translation of titles of books, tables of contents, captions of tables and graphs, and individual sentences and paragraphs. It has been successful and has virtually supplanted walk-in oral translations.

Raw machine translation made available to the analyst was followed, in 1993, by the use of raw machine translation for foreign materiel exploitation (FME). In the fall of 1990 a large cache of Soviet and East German documentation was made available to Germany, Britain, and the United Stated for exploitation. This turned out to be the largest collection of documentation made available to U.S. intelligence since the end of World War II. The original amount was 180,000 pages of Russian and German, but as new equipment and documentation were acquired, this number grew.

The volume, however was so great and the turnaround requirements so demanding that neither in-house resources nor contractor translators could keep up. Nellis Air Force Base, Nevada, also using SYSTRAN, enlisted engineers to type in the text and they initiated in 1992 the use of full-text raw machine translation. They had no linguists for editing and paid engineers to type in the material because they needed the translations so badly. Nellis AFB also initiated the first "side-by-side" hard copy, allowing the analyst to quickly reference a xerox copy of the original page, thus avoiding the necessity of cutting-and-pasting of graphics. In 1994 NAIC, faced with overwhelming translation requirements and loss of personnel, adapted the side-by-side approach, producing raw MT for most FME products. To date some 30,000 pages of Russian and German documentation have been translated this way.

Down With the Old, Up With the New

The end of the Cold War brought dramatic changes at NAIC. Except for the unexpected windfall of FME documents from Germany, the traditional emphasis on Soviet and Soviet Bloc intelligence dramatically decreased, with a gradual shift from 90 percent emphasis down to about 30 percent. NAIC now took on a global intelligence responsibility. This expanded coverage and the downsizing of the military establishment forced a new appraisal of open source intelligence as the source of first resort intelligence and large-scale information systems such as Internet. The requirement for efficient processing of large volumes of foreign literature gave new impetus to OCR and MT development.

MT Joint Ventures

One of the main outcomes of these changes was an increase in the variety of MT systems developed by the government and the collaborations established by different organizations in development. For example, Foreign Broadcast Information Service (FBIS) and NAIC had collaborated in developing Systran Japanese since 1985. In 1992 the Joint National Intelligence Development Staff (JNIDS) and NAIC joined together to develop a Spanish system. In 1994 NAIC and the FBI began collaboration in developing a Chinese system. NAIC independently has developed Korean, Portuguese, and Italian, while CIA/ORD began Arabic development in 1995. NAIC and CIA/ORD for several years have collaborated in developing Cyrillic, Arabic, and Chinese OCR'S.

Mainframe to PC

In 1992 the U.S. Government began funding the conversion of Systran from IBM Assembler to C/C++ language. The reasons were obvious: cost reduction for installation, the popularity of powerful PC's and workstations, ease in maintenance, ability to integrate MT with other software and last, but not least, nobody was coming out of universities trained in IBM Assembler, The Japanese system was the first one converted (FBIS funding), Xerox funded the conversion of English source systems, and NAIC funded the conversion of the remaining systems. In 1994 Systran was selected as a DoD Migration System. This meant that Systran would have to run in a client-server UNIX environment. Fortunately, most of the conversion had already been done before this DoD directive came out. All new systems developed at Systran are now done in C (Chinese, Korean) and the IBM VM systems will be phased out over time. Systran systems now run under DOS Windows and UNIX-based SunSparc computer environments.

Prospects Look Good

Given the revitalized interest in MT by the Intelligence Community and the Government at large, MT development looks promising in the future. Several key government organizations have joined together to back future developments. These include COSPO (Open Source Programming Office), CENDI (an interagency group composed of Commerce, Energy, NASA, National Library of Medicine, and NAIC), and the DODIIS Migration System panel. The following projects are envisioned for completion in the near term.

Systran Windows

Systran Windows versions falling under the purview of Government perpetual licenses will be made available to far-flung remote Government users, ranging from warfighter squadrons to small laboratories.

Open Source Information System (OSIS)

The OSIS is an Intelligence Community-sponsored network that provides connectivity to open source information for intelligence community users and other organizations interested in open source. By definition all material is unclassified. It is available through the Internet and six nodes exist today. By the end of 1995 there should be 19 nodes, with future anticipated connectivity to U.S. Unified and Specified Commands and selected U.S. embassies overseas. The OSIS makes databases, gray literature, libraries, open source requirements, toolkits, and processing capabilities available to users. Systran MT will also be made available. Systran German, French, and Spanish systems are already loaded and, in the future, interfaces with standard wordprocessing packages and OCR's will be incorporated.

Intelink

Intelink is a system put together by the Pentagon and CIA in December 1994. This is a world-wide computer network that has 35 intelligence organizations feeding it and so far more than 3,000 users, all with secret or top-secret security clearances to tap into the system. Intelink operates over the Pentagon's Defense Systems Network and its successor the Joint World-wide Intelligence Communications System. It has its own lines or leases special lines from phone companies to send encrypted message. Although Intelink mainly provides finished intelligence, some tools will be made available and Systran as a migration system will be loaded.

CIRC Revisited

During the Cold War NAIC developed a large storage and retrieval database of science and technology intelligence information and references to documents acquired. It was called CIRC. By 1993 there were some 10 million documents referenced, mostly covering the former Soviet Union, Eastern European countries, and China. But with the collapse of the former Soviet Union, downsizing, and in particular tight budgets, it become apparent that the Cold War largesse of 3.5 million dollars for populating and maintaining the database was untenable, and more automation had to be introduced in all phases - processing of data, indexing, search, and retrieval.

Expensive manual translations of extracts and manual indexing for CIRC will soon be replaced by a new way of doing business. Some of the key features include massive use of OCR's, MT, full-text translation, NLP, and SGML. The assumptions are that approximately one million pages of data will be collected annually! The majority of the processing must be done on hardcopy documents. Despite a worldwide move toward electronic publishing, about 90 percent of the world's data is still available only in hardcopy. The idea is to set up a bank of OCR's to scan the documents for MT and further processing. The government-sponsored Cuneiform (10

[Note: next page missing in original ‘MT Summit V’ proceedings.]

leading-edge technologies covering a wide spectrum of applications. It was this aggressive risk-oriented mindset that offered the climate to acquire, experiment with, and ultimately prove the cost-effectiveness and viability of MT. Several farsighted managers who saw the necessity of consistent funding made an immense contribution. Consistent funding and a can-do attitude - it cannot be overstated - were key factors in getting MT systems operational.

MT Culture

It was very significant for us to establish an MT culture in the translations department. Originally we were a human translations shop, then we were divided into two sections in the early 70's, a human translation group and a machine-aided translation group. With the advent of Systran, by 1975, MT had gained dominance. MT had saved the translations department by providing our big-ticket high volume product. MT enabled us to train translators as lexicographers and eventually to hire MT people instead of human translators. The evolution of our translation shop into an MT shop has taken a long time, and it has been painful, but ultimately it led to full support of MT by translators and full support of MT throughout the organization.

Feedback

Feedback by NAIC to Systran developers was important in advancing MT progress and gaining support of the systems by post-editors. At one time we had six people working in the so-called "Update Group" providing feedback to the Russian system. This was in addition to the Russian post-editors who constantly collected not-found words and spotted ridiculous translations. Moreover, with ten years of digitized data we were able to collect vast amounts of information to support enhancement of linguistic problems of all sorts. Lastly, we became very knowledgeable about the systems, came up with many of the ideas that became standard fare in all Systran systems, and ran all new versions through tens of thousands of random sentences in order to evaluate new releases for acceptance or rejection. A deep understanding of Systran and a doggedness in pursuing problems and increasing dictionaries was one of our trademarks that eventually spelled success.

The Customer is Always Right

Winning customer support of MT was critical. Our organization became a strong advocate of total quality management in the early 90's, but even before this we had used many of the basic tenants to win over customer support. First of all, we provided tailored MT products to analysts, although if nothing out of the ordinary was required, we had a default product called "partially-edited machine translation". But we were basically very solicitous to give customers what they wanted: raw machine translation, machine translation devoid of all organizational markings and edited to

human translation quality to protect sources, soft-copy MT, special formatting and sensitivity markings, or machine translation edited to high technical accuracy.

When initiating something new in MT we would provide briefings for analysts and we added a user questionnaire to all MT translations asking for feedback from subject-matter experts on terminology. But more than anything else we pressed to meet any user requirement on turnaround time because we ultimately knew that speed in getting the translation in the users' hands was the strongest selling point of MT.