Multilingual Processing for Operational Users

Keith J. Miller, Florence Reeder, Lynette Hirschman, David D. Palmer

The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7508
{keith, freeder, lynette, palmer}@mitre.org

Abstract

This paper describes multilingual technology projects currently being undertaken in conjunction with the NATO BICES (Battlefield Information Collection and Exploitation) organization. First, we describe the basis of the multilingual processing for these projects, the CyberTrans machine translation environment, an operational system that enables the use of machine translation (MT) by intelligence analysts [1]. We will briefly describe the impetus behind the development of CyberTrans as well as the system design and implementation. Next, we will discuss the operational pilot installation of CyberTrans on the BICES network. Finally, we will present some potential multilingual technology pilot experiments for BICES. These future technologies will enable NATO to meet the challenges of its inherently multilingual user community and will pave the way for interoperability across language barriers in the future.

1. Introduction

The challenge of automatic processing of language data from multiple languages is increasingly diverse and problematic. In addition to the wealth of work that has been done on the English language, "foreign language" processing needs are increasing, in part because of the changing conditions and needs in the world. Traditionally, users could focus on just a few foreign languages and a limited number of sources of foreign language materials. As we begin the 21st century, users of online materials are faced with having to process, utilize and exploit documents that may be in one of many languages or a combination of languages. It is not feasible to expect a given user to know all of the languages related to their topic of research. It is equally unrealistic to expect to have on-demand translators available in every language whenever they are needed. Because of the expanding need, tools are being developed to automate the use of foreign language materials.

A key component of many multilingual applications is machine translation (MT). A common vision for machine translation is as a small part of a larger process that is partly or completely automated. For many users, this does not mean having to work with yet another tool and yet another interface, but a nearly invisible companion that incorporates translation and necessary support technologies. One such system, the United States Army Research Lab (ARL) FALCon system [1,2], combines scanning, optical character recognition (OCR), translation and filtering into a single process. Another view of this is the DARPA Translingual Information Detection, Extraction and Summarization effort (TIDES) [3]. TIDES represents the pinnacle of information access and is a significant challenge for MT. MT supports the translingual aspects of the effort and here can be viewed as an embedded tool that facilitates other technologies. Finally, the integration of MT into the process for intelligence analysis serves as the basis for the CyberTrans project [4].

2. CyberTrans

The first incarnation of CyberTrans grew out of a demonstration that machine translation could be useful in the intelligence analysis process. As a result of a survey of MT technology (Benoit et al, 1991), it was believed that MT was ready for incorporation into an operational environment. Questions remained, however, about which commercial off the shelf (COTS) or US government off the shelf (GOTS) MT engines to use, and how to make them accessible in a user-friendly manner. Thus, CyberTrans itself is not machine translation per se, but it is a way to make machine translation (both COTS and GOTS) tools available to a wide range of linguists and analysts. It incorporates the Systran family of MT systems, which provides several language pairs (German, French, Spanish, Portuguese, Italian, Russian, Serbo-Croatian, Ukrainian, Chinese, Japanese and Korean to English) free to for US government use. In addition, CyberTrans incorporates the Globallink (Lernout and Hauspie) tools, which provide German, French, Spanish, Russian to English translation, as well as the GOTS product Gister. Initially, CyberTrans was designed as a wrapper around MT systems in Unix environments. Based on a client-server architecture, it provides a common user interface to its multiple translation engines. The server software interacts with the translation engines, controlling the flow of the translations, and the client software handles the user end of the transaction. Historically, four clients were provided: e-mail, web, FrameMaker, and command line. By providing translation through these media, users could translate documents in a familiar interface without having to be concerned with differences between translation products.

Shortly after the fielding of the initial prototype, the need for additional language services to accompany translation became apparent. The "real world" data sent to the translation engines pointed out the differences between translation in an interactive environment and translation in an embedded, automated environment. Interactive translation is much more forgiving of low quality input data while automated processing must handle issues arising from data quality on the fly. Given the assumption that the user has no control over the production of the source document, and hence no control over the quality of that document, a series of pre- and post-processing tools were incorporated into CyberTrans, thus transforming it from a user-interface wrapper to a

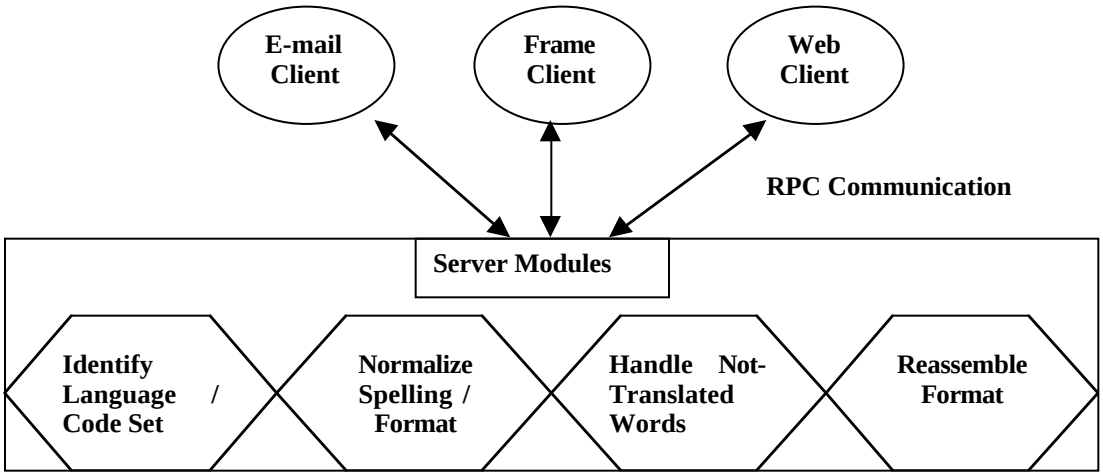

Figure 1: Original CyberTrans Architecture

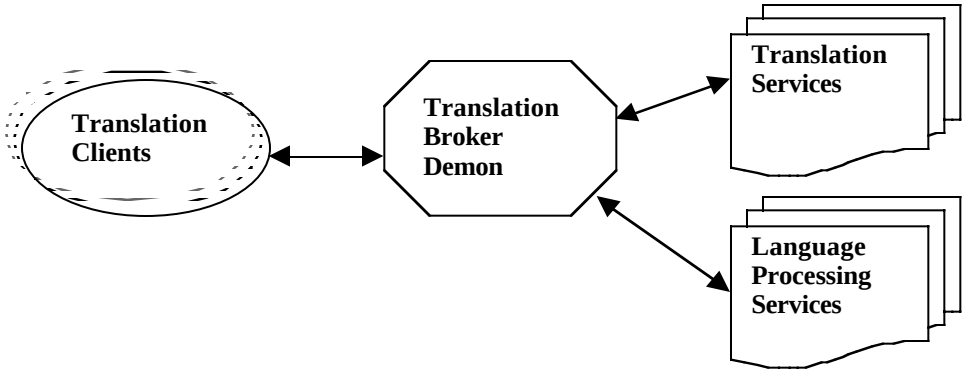

Figure 2: Updated CyberTrans Architecture

value-added machine translation environment. Initially, these pre- and post-processors were included in the functional architecture as depicted in Figure 1.

The server portion of the shell incorporates a) language and code set identification; b) language and code set conversion; c) limited spell checking (particularly diacritic reinsertion); d) format preservation. The data flow follows the following steps. Upon being submitted to CyberTrans, a document proceeds through steps a-d and is then passed on to the appropriate translation engine. The results are re-packaged, and the results are sent to the client for presentation to the user. As mentioned above, a number of clients are available: an e-mail client (send an e-mail to a specific address, get a translation back); a web-client (cut and paste or provide a URL); a command-line client and an API. The API allows integration of CyberTrans into a number of processes including word processing packages (FrameMaker; LotusNotes) and other applications, such as the TrIM, MITAP, and Open Sesame applications, discussed in Section 5.

The increased complexity caused by the addition of these language tools caused a necessary re-design of the architecture from a client-server model to an enterprise service model. This model is characterized by an open architecture of loosely coupled modules performing services for multiple applications. In this architecture, daemon processes broker translations. A request for translation is passed to the system by a client program, and a translation plan consisting of a series of translation-related services is created. Each service is requested from the responsible system object, and the resulting translation is passed back to the client programs. Implemented in a combination of C++ and Java, the new version represents a service-oriented architecture. Figure 2 shows this updated view of the architecture.

Language processing services include language/code set identification; code set conversion; data normalization, including diacritic reinsertion and generalized spell checking; format preservation for Hyper-Text Mark-up Language (HTML) documents; not-translated word preservation and logging, and others. The available clients remain e-mail, Web and FrameMaker. Platforms include both Unix and PC for clients, with the capability to incorporate PC-based translation tools as part of the service.

Many of the modifications and improvements in the system came as a direct result of the deployed of CyberTrans in an operational environment with a steadily-increasing user base. As can be seen in figure 3, between April 1998 and May 2000, CyberTrans experienced over 600% growth in monthly usage. In Section 3, we turn to

lessons learned as a result of having an operational MT capability, running 24 hours a day, 7 days a week,

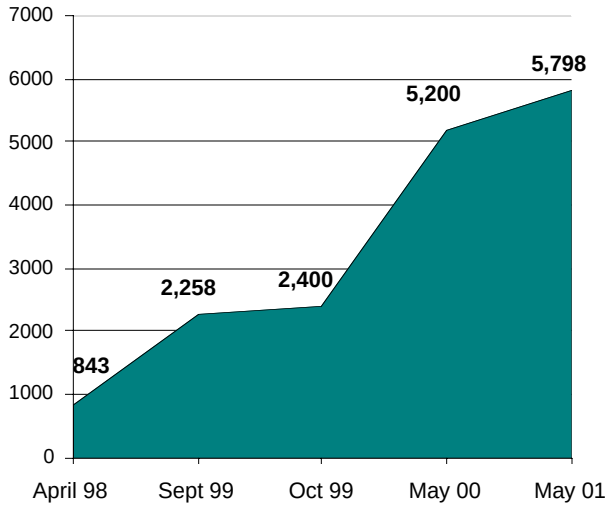

Figure 3 CyberTrans Usage Statistics

servicing over 4500 translation requests per month

3. Operational Machine Translation on BICES with CyberTrans

Since its initial installation, CyberTrans has been running around the clock at a US government site. It is currently available on a secure internal network and processes between 4500 and 6000 MT requests per month in up to 30 different languages. While not approaching the "Star Trek" vision of perfect MT quality, it does provide useful translations for the purpose of scanning and filtering by intelligence analysts. Since its initial deployment as an operational system, we have learned much in the realm of real world translation. One obvious lesson is that the better the quality of machine translation is, the more invisible translation is to the user. Since our users range from those who cannot identify the language of the document they wish to process to linguists trying to speed up their work, we are familiar with a wide range of issues pertaining to users with respect to MT language processing.

This success led to interest in CyberTrans from other organizations in which the US government is involved. In addition to the initial system deployment, we have recently installed a version of CyberTrans on the NATO BICES (Battlefield Information Collection and Exploitation) network. The BICES network facilitates the coordination of battlefield intelligence gathering among the NATO nations. BICES is an ideal candidate for a pilot CyberTrans for two reasons. First, although French and English remain the official operational languages, NATO has 19 member countries representing upwards of 15 languages and has much to gain from an operational installation of usable MT. Second, access to native speakers of so many languages is a rich source of feedback for continued improvement of the CyberTrans environment and its component technologies.

Following from these characteristics, the following goals were set forth for the pilot installation of CyberTrans on BICES:

- Provide a standing MT capability on BICES that can be used by BICES users via the BICES Backbone Network (BBN) and can be improved over time.
- Provide feedback from BICES users to MITRE's DARPA TIDES team throughout this process.

This installation was realized with the support of the DARPA TIDES program.

Because of the quality of the output produced by state-of-the-art MT systems, there must be a balance between "selling the technology" and managing (potential) users' expectations in operational environments. In short, MT is a useful technology if it is not oversold, but also not undersold. For the BICES installation in particular, the following questions were focused on:

- Who is the user or customer?
- What are the user's requirements?
- How does MT fit into the overall user's process?
- What is the price/risk of miscommunication?
- What are the user's expectations?

For BICES, the initial users are volunteer native speakers of the various languages supported by the MT pilot. These users provide valuable feedback as to the utility of CyberTrans and on opportunities for improvement. After this initial phase, CyberTrans is made available to a wider community of users, who primarily use CyberTrans for purposes of information assimilation, such as understanding documents published in a language other than their native language. It is hypothesized that a small number of users will also use CyberTrans to assist them in producing certain documents for which there is an English-language reporting requirement. In both cases, users access machine translation on an as-needed basis, from their desktops, to translate documents that they already have in electronic form. Given the users' understanding that this is a pilot application, the output of the translation is not expected to be perfect. Furthermore, with the same limitation in mind, it is expected that users be highly skeptical regarding any seemingly suspicious translation output, which will hopefully mitigate the risk of any miscommunication. Finally, the BICES support team and initial users are briefed on the prospects and limitations of MT technology in general, in an effort to manage users' expectations of the technology. This information is to be passed on to the end users of the CyberTrans application. Additionally, messages about the realistic use of machine translation are prominently displayed in CyberTrans' web page interface.

With these answers to the guiding questions in mind, CyberTrans was installed on the BICES application test facility in May 2001 and was given a limited release on the operational network (the BBN) in late June. In July, the application was fully released and could be accessed by all BICES users. This initial implementation includes translation from Russian, German, French, Spanish, Italian, and Portuguese into English. Users have enthusiastically been providing feedback bearing on the translation quality, the pre- and post-processing facilities, and on the user interface, all of which will be used to enhance future versions of the system. In addition, users of the Portuguese to English translation facility have been

particularly positive in their response and have requested the addition of translation in the reverse direction.

4. The Value of User Feedback: Tailoring CyberTrans

The ultimate success of automated language processing applications depends on the breadth, depth, and overall quality of the lexical information being used in the systems. Accordingly, the highest portion of the cost of providing an MT capability reflects the amount of lexicography – or domain specialization – that must be done. It can total up to 70% of the cost of an MT engine and represents the greatest source of user dissatisfaction. In addition, many applications require specialized lexical repositories that reflect unique domains such as military, legal, scientific and medical terminology. We must find ways to update lexicons intelligently, using sources such as dictionaries, working aids, specialized word lists and other information reservoirs to provide broad vocabulary coverage. Our principal current approach is to record the list of words that do not translate and automate the handling of these.

Now that CyberTrans is in the operational phase of its pilot installation, a process will be put in place whereby logs of not-translated words are transferred back to the development team for use in updating the MT lexicons. Since CyberTrans can incorporate various MT engines, it is an interesting problem that different translation engines encode lexical entries in different ways, such that sharing lexicon entries between translation capabilities is problematic. We are working on lexicon service bureau (LSB) research designed to facilitate the sharing of lexical materials. One part of this is the automatic extraction of lexical entries from on-line, machine-readable dictionaries. Another part is the analysis of not-translated words. Each advance in this realm increases the overall quality of the output produced by machine translation systems.

In addition to domain-specific jargon and acronyms, proper names ("named entities") represent a complication for translation. Users of the pilot installation of CyberTrans on BICES have remarked that this is a particular problem for the Spanish to English translation output. In a recent test conducted on 100 news articles translated from Spanish to English, 2600 named entities were found. Two translations produced by human translators agreed on the "proper" translation of names only about 90% of the time. This, along with the user feedback, is a further indication that this phenomenon needs to be studied more in order to determine how to best handle these proper names in MT.

5. What the Future May Hold

5.1. TIDES Tools

In addition to being used to produce translated documents as an end product, CyberTrans is also being utilized in the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program. Its role in TIDES is to support the goal of information access from a variety of sources in multiple languages. DARPA's TIDES program is working with NATO partners to integrate language processing and translation technology into future intelligence networks.

This effort is supported in part by an Integrated Feasibility Experiment led by The MITRE Corporation, with a focus on automatically filtering, extracting and summarizing information about the outbreak and spread of natural and man-made disease. This technology integration and application effort called the MITAP (MITRE Text and Audio Processing) System uses the CyberTrans embedded machine translation capability to translate information from languages such as Portuguese, Chinese, Russian, or Spanish into English. MT is a key component within TIDES research and has been a top requirement of the NATO member nations' priority requirements. In addition to the MT technology itself, the BICES technology team has expressed some interest in the other component technologies contained in the TIDES effort. It has been suggested that due to its multinational nature, the BICES network would be a rich operational testbed for some of the technologies that make up the "IDES" portion of TIDES.

5.2. TrIM

In coalition operations such as those supported by NATO, participants with a wide range of native languages must be able to coordinate their efforts. Collaboration between coalition partners currently relies on the ability to settle on a single common language for all participants. This arrangement creates communication bottlenecks, and is likely to work less well at lower echelons and in the field than it does at higher echelons or in command centers. Additionally, the information that must be communicated (including source documents in native languages) can be in multiple languages and specialized domains.

From a linguistic point of view, translingual collaborative environments represent a new and exciting area of research. From an operational standpoint, emerging technologies in translingual collaboration enable us to envision environments in which nations are able to collaborate across language barriers in ways never before thought possible. Translingual collaboration produces challenges that have heretofore not been addressed. Linguistic analysis for purposes of natural language processing applications has typically fallen into one of two camps: spoken interaction between two or more active participants and written interaction for information dissemination or assimilation. Each of these has unique characteristics, and there is often little overlap between the two. Collaborative environments, however, yield a new form of interaction, one which has some characteristics of spoken interaction and some characteristics of written interaction. This mix presents unique challenges for MT-mediated collaborative computing.

Plans are currently in the works for implementation of a Translingual Instant Messenger (TrIM) prototype on the BICES network. As part of this pilot effort, we will be able to collect data highlighting the multilingual challenges faced by collaborative environments, including the unique interaction style, the specialized needs of the interactions, the difficulty of analyzing actual language use, and the inherent difficulties of MT use in an interactive environment. These studies will enable the improvement and tailoring of MT technology for this specialized use; it is anticipated that such improvements will take place along several dimensions and will include data normalization, lexical improvements, and syntactic enhancements both in the analysis and generation phases of translation.

We have demonstrated the appeal of a Translingual Collaborative tool with Translingual Instant Messenger (TrIM) but this gives us only a view to the future, not the pieces necessary to make that future a useful reality. To make the translingual sharing of a reality, we need to develop tools for capturing, analyzing and enabling translingual information sharing. For instance, in TIDES we need a way to rapidly acquire domain-specific terminology and make it usable to the automated processing tools. In TrIM, the translation capability must be more reflective of the language style and terminology that is actually required for collaborative, coalition work.

5.3. Open SESAME

Open SESAME is the BICES intelligence production and discovery system. Key to sharing information between the nations is a "library card" representing product information in the form of metadata. This system is the bedrock for information exchange between the seventeen BICES nations. The BICES nations produce all their intelligence in their native language, which is far too much to reasonably translate manually. The library card concept allows nations to publish products in their native language, capturing the document metadata in English. Discovered documents deemed important may be passed on for translation.

The planned incorporation of CyberTrans with Open SESAME will carry this concept and process to a higher level. During the submission process, it will be possible to pass metadata fields, such as title and summary, through CyberTrans, automatically generating an English equivalent for the library card. Similarly, intelligence researchers may then submit native language title and summary searches, which will then be passed through CyberTrans, to query the Index Server's English metadata tags for relevant intelligence products. Finally, free texts documents, may subsequently be passed to CyberTrans for full machine translation.

6. Conclusion

Due to its multinational character, the NATO BICES agency is a natural proving ground for multilingual technologies. It is also one of the organizations that stands to gain the most from the successful implementation of such technologies and from improvements that result from live pilot installations such as the one described in this paper. Working in conjunction with the BICES agency, we can make valuable multilingual technologies available to those who need them the most, while at the same time gathering crucial data that will enable researchers and developers to push those technologies to the next level of performance.

7. References

[1] Voss, C. R., Van Ess-Dykema, C., "When is an Embedded MT System "Good Enough" for Filtering?", *Proceedings of Embedded Machine Translation Systems - Workshop II, ANLP-NAACL2000*, Seattle, May 2000.

[2] http://rpstl.arl.mil/ISB/falcon.htm

[3] Hirschman, L., Concepcion, K., Damianos, L., Day, D., Delmore, J., Ferro, L., Griffith, J., Henderson, J., Kurtz, J., Mani, I., Mardis, S., McEntee, T., Miller, K., Nunan, B., Ponte, J., Reeder, F., Wellner, B., Wilson, G., Yeh, A., "Integrated Feasibility Experiment for Bio-Security: IFE-Bio A TIDES Demonstration", *Proceedings of HLT 2001 Human Language Technology Conference*, James Allan, ed., San Diego, March 2001.

[4] Reeder, F., "At Your Service: Embedded MT as a Service", *Proceedings of Embedded Machine Translation Systems - Workshop II, ANLP-NAACL2000*, Seattle, May 2000.