MT in the Online Environment: Challenges and Opportunities

Mary Flanagan
(mflanagan@csi.compuserve.com)
CompuServe Natural Language Technologies

Abstract

Integrating machine translation in online services poses a unique set of challenges and opportunities. Through the information superhighway, speakers of different languages can now communicate without the obstacles of time and distance. Yet language remains a barrier to communication in many cases. Draft quality machine translation (MT) can serve as a communication facilitator among online users who speak different languages. For the first time in the history of MT, millions of users will be exposed to machine translated text through online services. The potential exists for MT to become a part of daily communication, rather than a tool of the specialist.

At the same time, the challenges of machine translating online text are formidable. Online texts are often hastily composed and contain spelling, grammar, punctuation and accent errors. The traditional strategy of using subject-specific dictionaries according to the text's topic is often ineffective in the online translation environment, since message topics are varied and highly changeable. Though fast by comparison to human translators, many MT systems will need to become a great deal faster to cope with the enormous volumes of online text, and the users' demands for near real-time information.

CompuServe has recently begun integration of machine translation in some of its online services. One of these services, the World Community forum, is an online discussion area where CompuServe members can converse on topics of cross-cultural interest. There are over 600 special interest forums available on CompuServe. Translations of messages are provided between English and French, German and Spanish. In its first month of operation, more than 900,000 words were translated on the World Community forum alone. CompuServe will also introduce an e-mail translation service in mid 1995. This service will provide low cost document translation at two quality levels, and will support major word-processing formats. Unedited MT output will be returned within minutes; most edited jobs within 24 hours at low cost to the user.

MT in the Online Environment

Opportunities

The integration of machine translation technology in the online environment offers enormous opportunities to the MT industry. Through online services, millions of computer users will be exposed to machine-translated texts, potentially establishing a high profile for MT within the software industry and promoting increased sales and development of MT systems.

MT can be an effective tool for facilitation of multilingual online discussions, allowing users who speak different languages to communicate with one another, and promoting the globalization of the information superhighway.

With the rapid growth of the online industry, MT is poised for a surge in public interest. The MT industry, as well as the consumers of online services can benefit; but only if MT software is effectively adapted to the unique translation characteristics of the online environment.

Challenges

Most MT systems were not designed to translate online text. Commercial MT developers have generally built their systems around a characteristic profile of user needs. The intended end user is a bilingual or a professional translator who needs to enhance translation productivity through the use of MT. The user is likely to be knowledgeable about the MT system's dictionaries and their effect on translation algorithms. The source texts are primarily from a well-defined subject domain. They are often well-written, possibly prepared by a small number of trained technical writers. The user's workflow can be organized to run translations overnight with postediting of texts the following day. The postedited output will be used for publication or where high-quality translations are required. In this model, the MT system is a productivity tool of the professional.

Translating for the online environment requires a complete redrawing of this profile. The user of online machine translated texts is most often a monolingual who has no experience with MT or foreign languages. The user needs near real-time translations, suitable in quality for information scanning. The subject matter of the texts for translation is highly variable, and is likely to drift even within a single message. The source texts are often hastily composed, containing erroneous punctuation, usage and spelling. Sentences are frequently incomplete, and interrogatives and personal and product names have high frequency. Writing styles can vary tremendously, since any member of an online service, as well as outsiders can contribute online text. Postediting of text is prohibitively slow and costly. In the online MT model, the MT system is a tool for rapid, understandable quality information assimilation by untrained users.

To be successful in the online environment, MT systems must produce extremely high speed translations, and usable quality raw output. Pre-editing functions to screen for personal names, company names, emoticons, commands and samples of source code will enhance output quality, as will better handling of interrogatives. Systems must be robust enough to handle source texts with punctuation and capitalization errors as well as a wide variety of writing styles and topics. Finally, MT must be integrated seamlessly into the online environment. Most online users get connected to obtain information and participate in discussions. They wish to benefit from online translations, but very few are interested in learning about the workings of the MT system or how to use MT most effectively.

A Portfolio of CompuServe's MT-based Services

The World Community Forum

Forums® are online discussion areas on CompuServe where members can post and read messages from a message board, access files from online libraries and participate in live conferences with other members. CompuServe offers hundreds of special interest and professional forums.

The World Community Forum, launched in February 1995, is a meeting place and discussion area for topics of international and cross-cultural interest. World Community members discuss world politics, trade, language, sports, cooking, and cultural traditions. They ask for travel advice, find penpals and get help with language learning.

Separate, parallel copies of the World Community Forum are maintained in English, French, German and Spanish, allowing multilingual communication between members who speak different languages. An English-speaking member, for example, would log onto the English version of World Community to read and post messages in English. If this member posts a message in English while on the forum, the message will be automatically collected by the machine translation service, translated to French, Spanish and German and posted to the appropriate target language forums at intervals of 3 minutes. While connected to the English World Community forum, the CompuServe member can read English messages, as well as messages which may have originated in French, German or Spanish but have been translated to English by the machine translation service. The section and message thread names, as well as the order of the messages are maintained in the same order on all four of the forums making it easy to navigate through any of the forums to see the translation of a message. In fact, every translated message contains a copy of the original message following the translation. Because fully-automated MT can produce only draft translations, it is often useful to have a copy of the source message as a reference.

Approximately 30,000 words are machine translated each day on the World Community Forum. CompuServe members can GO WCOMMUNITY to visit the World Community Forum.

The MACCIM Support Forum

The MACCIM Support Forum provides software support to users of CompuServe's MACCIM online navigational software. Machine translation between English and German or French was integrated into the forum in August 1994 using the parallel forums model. Users of the forum post messages to inquire about problems using MACCIM, provide suggestions to other users, and learn tips for using MACCIM effectively.

Approximately 28,000 words per day are machine translated on the MACCIM Support Forum. Members can GO MCIMSU to visit the MACCIM Support Forum.

CompuServe Document Translation

CompuServe will introduce a document translation service in summer 1995. The service will allow users to upload files for rapid, machine translation between English and French, German or Spanish. Two quality levels will be available: unedited MT output, and postedited output. Most unedited translation jobs will be returned to the user in minutes; postedited output will be returned within 24 hours in most cases. The service will provide translations at extremely low cost through a convenient, easy-to-use interface.

Behind the Scenes

MT on Forums

The forums MT model maintains separate but parallel copies of the forums in English, French, German and Spanish. The section and message thread names, as well as the order of the messages are maintained in the same order on all four of the forums making it easy to navigate through any of the forums to see the translation of a message. Each translated message contains a copy of the original message following the translation. Because fully-automated MT can produce only draft translations, it is often useful to have a copy of the source message as a reference.

CompuServe's translation process is conducted at 3-minute intervals. First, a forum management software identifies and collects new messages. The messages are then submitted to a pre-editing software which identifies personal names, company names, emoticons, commands and samples of source code which may occur in the message. Translating these items can produce undesirable results. The pre-editing software marks the strings to prevent translation and to affect their rearrangement in the target language.

The pre-edited messages are then submitted to the translation software. Translations are performed by Intergraph's Transcend software, running on a Windows NT PC. The translation process is completed in near real-time with processing speeds of over 3,000 words per minute.

The Transcend software supports translation between English and French, English and German, and English and Spanish, however foreign-to-foreign (e.g. Spanish to French) language pairs are not yet available. To remedy gap in language pair coverage, CompuServe's translation process posts English translations of messages which originated in French, Spanish or German on all four of the forums.

Finally, the translated output is returned to the forum management software for posting on the appropriate target-language forum.

The CompuServe Natural Language Technologies Group has developed specialized dictionaries for many of the message sections in the World Community Forum. These dictionaries are automatically loaded for use with messages from specific message sections. We currently have available dictionaries for Food and Beverages, Travel, Trade, Politics, Recreation, Online language and Computer Science. In addition, a dictionary containing MACCIM terms is used for the MACCIM Support Forum. Dictionaries presently range in size from 500-3000 words, and are in continual development.

Document Translation

CompuServe's document translation service automatically uploads the user’s text prompting for information about language pair and postediting. Users can select unedited, “raw” MT output, or a postediting service. A central scheduling program manages the flow of incoming and outgoing documents, routing them to the translation software, collecting the translated texts and forwarding them either to the sender or to the postediting service if the user has selected postediting. Additional programs have been developed to calculate and post charges to the user's account, and load the appropriate dictionaries based on language pair.

The CompuServe Information Service

CompuServe is the world’s largest information service with over 3 million members worldwide. Membership in Europe, Asia and Canada is growing rapidly, and the integration of machine translation technology is one component of CompuServe’s effort to localize its online offerings. Members can access the CompuServe Information Service via a local phone call using CompuServe’s own network in most U.S. cities and 52 locations outside the U.S. In addition, members can access CompuServe outside these locations from nearly 140 countries via gateway networks. CompuServe provides English, French, German and Spanish versions of its CompuServe Information Manager (CIM®) software.

Soon to Come

CompuServe plans to introduce MT capability into its File Finder Service in late 1995. The File Finder allows users to search for online files containing text, images or software. Users can enter the filename, submission date or keywords to initiate a search. A keyword search is particularly difficult to conduct in a foreign language, because it requires the user to think of relevant keywords in an unfamiliar language. Machine translating the File Finders to French, German and Spanish will allow non-English-speaking CompuServe members to search online for files in their own language.