Machine Translation in an Information Management System

Seong Yong Kim, Robert A. Wisher,
Jon H. Brundage and Gwang Ryeol Jung

The purpose of this paper is to report on the development and testing of an English-Hangul, Hangul-English machine translation system. The system, known as the Hangul English Support System (HESS), is part of the Theater Automated Command and Control Information Management System (TACCIMS). TACCIMS is an extensive bilingual information network of over 400 workstations located throughout the Republic of Korea and Okinawa for use by the Commander-in-Chief Combined Forces Command (CINC CFC) (US/ROK). The HESS machine translation component plays a central role in translating a variety of input types, such as word processing files, data base requests, map graphics, and electronic mail. HESS is an important contribution to machine translation technology because: (1) It has a large scope of translation responsibilities; (2) It operates in a real-time, operational environment; and (3) It is designed for use by monolingual users. The paper describes the linguistic approach to translation that is based on lexical functional grammar and chart parsing. The development of the 60,000+dictionary entries is described. An interactive pre-editing mode is presented as an on-line tool for resolving lexical and syntactic ambiguities. The man-machine interface technology that underlies the pre-editing feature is discussed later in this paper. The paper also describes the development of a strategy to test the accuracy of the translations, focusing on the development of test sentences in the source text and measuring accuracy of the target text through an edit ratio function. Finally, the paper considers approaches to enhancing the overall performance of machine translation technology for information management systems of the future.

Introduction

Information management systems provide a real-time information-base for problems confronting an organization, furnishing timely operational and administrative information to decision makers. These systems represent a centralized means for planning, managing, and controlling an operation, be it a unclear power plant, an air traffic control system, or a large military organization. At the center of such a system are the individual users who must update databases, develop information packages, perform analytical tasks, and communicate to other users as part of their everyday work.

The enlarging scope of on-line information management systems is imposing new demands on integrated computer technologies. In particular, the increased interoperability between computers and software packages is a continuing concern in the development of a state-of-the-art system. In multinational settings, the management of information encoded in different natural languages presents an added dimension to the interoperability concern — the exchange of text between users who speak different languages. The growing cooperation between nations in the pursuit of common goals is likely to extend the requirement for a bilingual component in command and control and information management systems. Users must be able to communicate readily between different languages.

A bilingual command and control system is one which, upon request, can have narrative text translated from one language (source text) to another (target text). This can, of course, be accomplished through a staff of bilingual users, each responsible for the translation of their input. Another approach is routing the source text through a highly qualified translator who translates source text manually and transmits the target text back to the requester. These approaches have obvious drawbacks. Requiring bilingual users sets rigid limitations on personnel selection; a translation bureau adds time and costs that can delay the forwarding of essential information to decision makers. An emerging alternative is to develop machine translation algorithms that can provide rapid, first-cut translations, and allow the user to assess the applicability of this information to a problem. If greater translation accuracy is needed, a human-aided machine translation can then be applied to add coherence and precision to the first-cut translations.

This paper focuses on the functional requirements for a bilingual information management system, and, by way of example, parallels the development of the TACCIMS system for Korean and English-speaking users.

Taccims Background

The TACCIMS system will perform vital military command and control functions for the CINC CFC during armistice, crises, exercises, and the prosecution of war. Because of the bilingual features throughout the TACCIMS system, both the ROK and US personnel utilizing the TACCIMS system will be able to share crucial military command and intelligence information among the different staffs and forces. This ability to commonly share information in a combined military force structure is the significant accomplishment of the TACCIMS system. The manner in which TACCIMS supports the combined ROK and US force to mutually defend the Republic will undoubtedly have tremendous application in other combined military operations throughout the world.

The TACCIMS network will securely link various subordinate commands of CFC to three operational headquarters locations. TACCIMS will be deployed to fifteen sites in the ROK and to one site in Okinawa. TACCIMS will be comprised of some four hundred Goldstar Personal Computers(PC-386). These four hundred computers function as workstations and fileservers in a homogeneous network. Each site connects the workstations via a Goldstar dual ringed fiber optic Local Area Network (LAN) that is interconnected to other sites via communications circuits provided by either the Korea Telecommunications Authority or the U. S. Government Defense Communications System. The LANs utilize the IEEE standard 802.3, Ethernet ; while the sites are interconnected via the X. 25 protocol.

The TACCIMS utilizes many commercial and internationally accepted automation (i.e., ISO) and communications standards throughout the system. This will support future connections, and provide a capability for future growth and diversity. The TACCIMS software capability is provided by hosting bilingual relational database management system, bilingual electronic mail (e-mail), bilingual Office Automation, bilingual business and digital map graphics software applications on the Goldstar workstations and fileservers. The operating system use to control these bilingual applications is UNIX System V version 3.2. This version is also bilingual in that the TACCIMS must be capable of handling the two coding standards for machine representation. These two standards, ASCII and Korea Standard Code (KSC) 5601, exist simultaneously on the system in order to accommodate the features required of a bilingual Korean-English system. This dual existence of standards makes the implementation of TACCIMS more than a trivial integration of commercial software and the hardware. A key feature of the TACCIMS is its ability to provide a machine translation capability between the Korean and English languages.

In the current architecture for the TACCIMS, translations can occur at three levels : workstation, fileserver, and Central HESS. The difference is in the size of the dictionaries and rule sets for translation resident at each level, and in the updates and software maintenance. The Central HESS is the most current, and is responsible for updates throughout the TACCIMS network.

Machine Translation Background

The early approaches in machine translation were oriented to word-to-word and phrase-to-phrase conversions. Here the computer served simply as a dictionary look-up tool (Slowm (1985)). These approaches produced translations of poor quality, largely because they ignored a more complete view of language in terms of linguistic components. Machine translation has advanced steadily over the past decade, largely due to the influence from the integrated fields of psychology, linguistics, and computer science. This interdisciplinary domain, known as cognitive science, has invited machine translation approaches to factor in a much deeper understanding of syntax and semantics (Carbonell (1981)). These cognitive underpinnings are now central to state-of-the-art research systems, and are likely to be the key to the long hoped for system that might someday produce highly comprehensible and accurate translations of unrestricted text.

Although this is a distant goal, substantial progress has been made in machine translation that makes possible its incorporation into on-line systems. In a restricted domain, such as has been demonstrated in an English-French system for translating Canadian weather forecasts, performance is excellent (Slocum (1985)). In broader domains where dictionary entries are extensive and topics are unrestricted, performance cannot be expected to be anywhere near 100% accurate. The present day advantages that machine translation offers are rapid and acceptable first-cut translations,. accurate conversions of nominal items contained in data records, and interactive tools to reduce the time required for more precise translations. Continuing interest and investment in these systems will over time lead to incremental improvements in translation performance.

In considering the development of a bilingual information management system, one must determine the breadth of the domain and the accuracy requirements. Restricted domains, such as weather forecasts, might be more conductive to a knowledge-based translation system which can depend on its represented domain knowledge to aid in the analysis of semantics and the selection of words. When dealing with a large-scale command and control environment where electronic mail and messages can address a wide range or topics, the construction of a knowledge base becomes impractical. Here, a more general, but less accurate, approach is necessary.

Functional Requirements

The task of machine translation is to accept the source language input and, retaining format and header information, translate this source into the target language. The computational approach to this process can vary between various direct and indirect methods (Pentheroudakis (1990)). Regardless of the method, there needs to be a framework to the flow of text, from an initial pre-editing process to an optional post-editing of text into a finished product ready for distribution. Depending on the need for precision, the degree to which each of these processes is applied will vary.

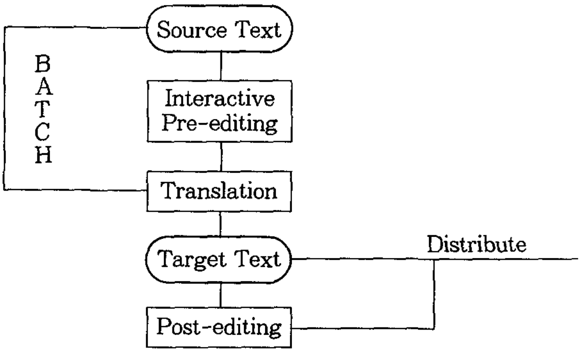

Figure 1. Principal Features of Machine Translation.

Figure 1 illustrates the principal operational features of a machine translation system. Each will be described in brief detail below.

Source Text

The source of text for translation is submitted to the machine translation system from various origins, depending on the environment. In our experience in a military command and control system, text can originate from the following information management sources:

- Word Processor
    - Reports
    - Briefings
- Electronic mail
- Database query
    - Data records
    - Spreadsheets
- Business and Map graphics
- Military messages

Batch Mode

Submitting a request by batch will bypass the interactive pre-editing process, with a consequential reduction in accuracy. The reduction is due to a word selection method for disambiguation that is based on frequency of use and syntactic category rather than the intended meaning. An example is provided in a later section.

Interactive Pre-editing

This optional feature behaves much like a spelling checker, in that it identifies words, phrases, or acronyms in the source that need to be clarified by the originator. The pre-editing resolves part of speech and other ambiguities, such as an intended meaning for a particular word. It is not uncommon for a word to have multiple meanings (word, sense ambiguity), such as the word "spring" which can mean a jump (도약), an elastic device (용수철), a season of the year (봄), a small stream (샘), or a recoil force (탄

력). Since each meaning corresponds to a different word in the targt language, the pre-editing forces the user to specify the intended meaning. The translation algorithm itself is not necessarily geared to accuracy in selecting alternative meanings, as it depends on statistical distributions.

The interactive pre-editor is also capable of resolving a syntactic ambiguity. Here, the alternative parses are represented by a bracketing scheme which distinguishes various noun phrase and verb phrase combinations. The user selects from the interface menu the corrcet alternative. Obviously, the pre-editor is closely associated with the translation engine, and serves as a window and decision aid to the translation process.

Translation

This process parses the source text into constituents, constructs an intermediate representation, and finally "transfers" this form to the target language. There are numerous ways in which the intermediate representation can be constructed and transferred. Two critical factors are the size of the dictionary and completeness of the rule set that drives the parser. Having a translation algorithm that is sensitive to the type of input is advantageous. By knowing that the source input is a label from a spreadsheet column, for example, the translation can fall back to a simple word or phrase conversion. A more striking example might be in the attempted translation of a viewgraph chart, which often conveys noun strings in a bullet format. Here, if the parser expects to translate a complete sentence, the lack of a verb phrase in a noun string will lead to a failed parse. By taking into account special conventions of language use, a sensitive translations algorithm can lead to more acceptable translations of a variety of inputs.

The translation engine for the Hangul English Support System in TACCCIMS is derived from lexical functional grammar (Bresnan (1982), Frey (1985)). Lexical functional grammar is a representation of the associations between the semantic arguments of sentences and their surface constituent structures. These grammatical relations are represented as f-structures in lexical functional grammar. An example of an f-structure for the following sentence is as follows:

"The sergeant entered the data."
"그 하사관이 그 자료를 입력했다."

```
┌ SUBJ   ┌ PRED     'SERGEANT'
│        │ SPEC     THE
│        └ ARG     ┌ NUM      SG
│                  └ PERS      3
│
│ PRED    'ENTER  <(SUBJ), (OBJ)>'
│ TENSE    PAST
│
│ OBJ    ┌ PRED     'DATA'
│        │ SPEC     THE
│        └ ARG     ┌ NUM     PL
└                  └ PERS     3
```

Translation Example

The following translation example illustrates the use and advantage of the interactive pre-editor.

"The data has been entered into the computer in the office."
"그 자료가 그 사무실에서 그 컴퓨터 안으로 들어가고 있었다."

This sample sentence was translated by the HESS in the batch mode. This particular example has two errors. The first is in the selection of the postposition for the word "에서." The correct translation should be "에 있는." This error was probably due to a transfer error. On the other hand, the selection of the word "들어가" (meaning to enter a room) was based on a statistical frequency rather than the intended meaning. These errors can be reduced through the use of the interactive pre-edit function. Here, the alternative meanings are presented to the user who, through a man-machine interface, selects the correct intention. In this case, the word "enter" would correctly translate to "입력되."

Target Text

This is the first-cut product of the translation process. Besides the caliber of the translation algorithm, the quality is a function of the complexity of the sentences in the source, the application of pre-editing, and the degree to which the user needs to understand details. Enhancements to the first-cut product can be made through post-editing.

Post-editing

A qualified translator interacts with the target test to correct grammar, re-word awkward phrases, fix semantic errors, and add overall coherence to the text. Post-editing makes a dramatic difference in the overall comprehensibility of text, and is essential for a document that requires exactness in its translation. In the HESS system, post-editing is accomplished through a "Linguist Mode" of operation in which the source and target are paired in an interleaving fashion. The post-editor applies word processing utilities to make changes in the target text.

Dictionaries

Dictionaries should include all likely words, phrases, acronyms, and idiomatic expressions that are likely to occur in a source text. In the HESS system, a range of up to 70,000 entries covers most occurrences. Based on frequency of usage, there will probably be around 5,000 words that represent about 90% of word occurrences. An analysis of this relationship for the Korean language (Ministry of Education (1956)) is shown in figure 2.

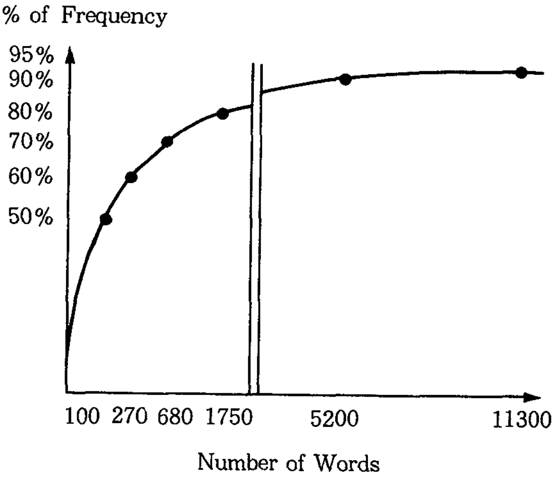

Figure 2. Relationship between Dictionary Size and its Coverage

The dictionary should be tied into the text preparation process. This would allow the identification of words that are not in the dictionary prior to a translation. Tracking word usage patterns is important if one is considering a distributed architecture which limits the size of the dictionary resident on a workstation. The most frequently occurring words should reside on the workstations.

The dictionary entries for the TACCIMS were selected from common usage dictionaries, military acronym lists, and an analysis of text provided by the translation branch of the CFC. A dictionary entry/edit tool assisted in the entry of each word. As there will inevitably be additions to the dictionary, dictionary maintenance is a recurring task. The dictionaries should be maintained by a linguist fluent in the language pair. A software utility that manages this function is an absolute necessity. In the TACCIMS, a useful source of new words is the stochastic log of those words which the translation algorithm was unable to translate. Another source could be notes that the user forwards during the interactive pre-editing. These notes can identify intended meanings that were not available in the selection menu for ambiguous words.

Performance Testing

Translation Speed

Just as wordy, complex sentences are difficult for a reader to understand, so too are they difficult for a computer to translate. Knowing that a document you are preparing will be submitted to a machine translation system, keep in mind a maxim for clear writing — keep sentences short and simple. Sentence length imposes on the parsing component of a translator, which must consider many possible paths when dealing with a lengthy sentence. There is an exponential relationship between sentence length and translation speed. This relationship is depicted in figure 3. Short sentences can be translated within a half minute, but sentences exceeding 15 words may require several minutes. A useful benchmark is a 6 word sentence in 6 seconds.

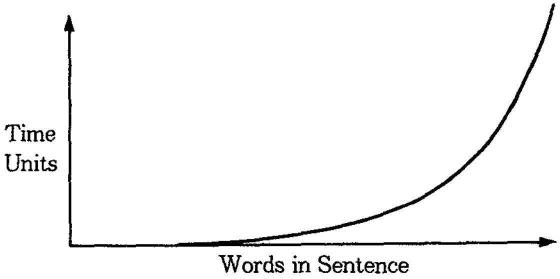

Figure 3. Relationship between Sentence Length and Translation Time

Translation Accuracy

Accuracy in translation is usually measured in terms of an edit ratio (ER): the proportion of words in an initial translation that are changed after post-editing (Guida & Mauri (1986)). Generally an ER of 20% or lower is considered quite good. If there is a need to have precise translations of certain text, such as a regulation, a linguist skilled in the art of translations will be required to perform post-editing. If on the other hand one needs only to understand the gist of a document, the first-cut translation offered by a machine translation algorithm can usually serve the purpose. For example, a decision maker might be able to discard a document based simply on its topic—a report on fuel supply is not immediately pertinent to a pending decision on staff reductions. Although the target text on fuel supply might have been from only a first-cut translation, its irrelevance to the decision at hand avoids the need for further precision. The much greater challenge comes from the need to have a quickly available translation that is reasonably accurate.

The application of an objective measure of ER is a key to a reliable indicator of how well a machine translation system performs. An example of ER is as follows :

$$ER = \frac{\Sigma\, \mu(Ti) * \rho(Ti)}{\Sigma\, \rho(Ti)}$$

$T = \{T1, T2, \cdots, Tn \mid Ti = \text{i-th sentence}\}$

μ = value of distance between correct meaning and meaning in target

ρ = importance function of set T

Where:

$\mu(Ti) = 0$:	correct
$= 0.2$:	word deletion
$= 0.4$:	word insertion
$= 0.6$:	word change
$= 0.8$:	word ordering
$= 1.0$:	unacceptable

The importance function ρ and the ER can have values ranging from zero to one.

Testing Performance

The ultimate test of performance will be the usefulness of the target language text to the users. In the course of development, the system should be tested with a suite of carefully constructed sentences that exercise the parser in systematic ways. Some research is ongoing on the establishment of appropriate test suites. However, not every measure of success is applicable to every system (Flickinger et al. (1987)). One approach in the TACCIMS is to develop a simple grammar capable of generating simple subject-verb-object sentences, with increasing complexity of each of these constituents. For example, the subject can be "The soldier," or "The American soldier," or "The American soldier near the computer," etc. Similar grammatical complexities can be added to the verb and object constituents. By varying the combinations, and then tracking the speed and accuracy performance by the machine translation algorithm against these various combinations, one can identify possible areas for improvement in the parsing algorithm.

Such a suite of over 700 test sentences have been develped for the TACCIMS system. An ER will be applied to each sentence, and speed will be clocked as a means of determining system performance. These sentences can then serve as a baseline of machine translation performance in tracking the extensibility of the system for meeting the final operational capability.

Future Considerations

Improving translation speed and accuracy will be the focus of future enhancements to real-time machine translation systems. Of course, speed improvements can occur simply through the use of faster machines. The more critical area of accuracy should make steady improvement with ongoing research in the area of computational linguistics. A near term possibility for improving accuracy involves the incorporation of a grammar-sensitive word processor into the office automation software. This utility would essentially pre-parse a sentence in an on-line mode. With the capability to detect immediately difficult text for later translation, adjustments can be made in the source language that will facilitate a forthcoming translation. A grammar-sensitive word processor could also reduce the workload during pre-editing, as the number of ambiguous constructs would be reduced. The extensibility of machine to achieve accuracies of 80% or greater is likely within the next decade.

Conclusion

Machine translation in an information management environment is a practical advantage for future systems. The example presented from the Hangul English Support System is a prototypical design for bilingual information management systems of the future. It is important to maintain realistic expectations as to how complete and accurate a particular translation can be, given the state-of-the-art of machine translation technology. Nevertheless, the introduction of this technology into information management environments is a exciting step forward.

References

Bresnan, J. (ed.) (1982) *The Mental Representation of Grammatical Relations*, Cambridge, Mass: MIT Press.

Carbonell, J. G., R. E. Cullingford and A. G. Gershman (1981) 'Steps towards Knowledge-based Machine Translation,' *IEEE Transaction on Patterns Analysis and Machine Intelligence*, PAME-3(4).

Flickinger, D., J. Nerbonne, I. Sag and T. Wasow (1987) *Toward Evaluation of NLP Systems*, Hewlett-Packard Laboratories, Palo Alto, CA.

Frey, W. (1985) 'Noun Phrases in Lexical Functional Grammar,' *International Workshop on Natural Language Understanding and Logic Programming*, Elsevier Science Publisher B. V.

Guida, G. and G. Mauri (1986) 'Evaluation of Natural Language Processing System: Issues and Approaches,' *Proceedings of the IEEE*, 74(7).

Ministry of Education (1956) *An investigation of Word Frequencies for Korean Language* (우리말 말수 사용의 찾기 조사), Seoul: Taesung Press.

Pentheroudakis, J. E. (1990) 'You Can Get There from Here: Design and Implementation Issues in Machine Translation System,' *Proceedings of the Topical Meeting on Advances in Human Factors Research on Man/Computer Interaction: Nuclear and Beyond*, American Nuclear Society.

Slocum, J. (1985) 'A Survey of Machine Translation: Its History, Current Status and Future Prospects,' *Computational Linguistics*, 11(1).

Dr. Seong Yong Kim / Dr. Gwang Ryeol Jung
Korea Institute for Defense Analyses
P. O. Box 250 Cheongryang
Dongdaemun-gu, Seoul
Korea

Dr. Robert A. Wisher
U. S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333
U. S. A.

Dr. Jon H. Brundage
Department of the Army
Project Manager TACCIMS
APO San Francisco 96301
U. S. A.