[Proceedings of the *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, Hamilton, New York, August 14-16, 1985]

New Approaches to Machine Translation

Jaime G. Carbonell and Masaru Tomita

Carnegie-Mellon University

Pittsburgh, PA 15213

3 July 1985

Abstract

The current resurgence of interest in machine translation is partially attributable to the emergence of a variety of new paradigms, ranging from better translation aids and improved pre and post-editing methods, to highly interactive approaches and fully automated knowledge-based systems. This paper discusses each basic approach and provides some comparative analysis. It is argued that both interactive and knowledge based systems offer considerable promise to remedy the deficiencies of the earlier, more *ad-hoc* post-editing approaches,

1. A Historical Perspective

Researchers in machine translation have aspired for three decades to develop highly-accurate, practically-useful, fully-automated translation systems. This ultimate objective remains as elusive today as it was in the late 1950's, although the field has seen considerable progress ranging from theoretical advances in computational linguistics to useful partially-automated translation systems. In the early heyday of machine translation, the rallying cry was *"95% accurate, fully automatic high quality translation!"* [13, 2]. In fact, that motto was repeated so often than it became an acronym: "95% *FAHQT".* However, little attention was paid to fundamental issues such as: exactly what does "high quality translation" signify?; what does it mean for a translation to be "95% accurate"? And, most importantly, little thought was given to the requisite theoretical underpinnings -- linguistic and computational -- that must be established and understood before fruitful system engineering can begin.

As discussed in [4], there are multiple dimensions of "quality" in the translation process, to wit:

- *Semantic invariance* -- Preserving invariant the meaning of the source text as it is transformed into the target text.
- *Pragmatic invariance* -- Preserving the implicit intent or illocutionary force of an utterance. The manner in which a proposition is stated may convey intent, urgency, politeness, etc. And, the translated text should convey the same implicit information to preserve pragmatic invariance.
- *Structural invariance* -- Preserving as far as possible the syntactic structure of the text under translation.
- *Lexical invariance* -- Preserving a one-to-one mapping of words or phrases from source to target texts.
- *Spatial invariance* -- Preserving the external characteristics of the text, such as its length, location on the page, etc.

Whereas early MT systems sought to preserve lexical invariance in the hope that all other invariances would follow, modern approaches take a somewhat more realistic view. Semantic invariance, for instance, is becoming a more dominant criterion -- with other invariances preserved only in the service of conveying the appropriate meaning. Given this criterion for accuracy, the motto 95% *FAHQT* rings rather hollow. First, it doesn't address the severity of the 5% errors -- are they simply misinterpreted nuances, or can they completely change the meaning and intent of the text? Second, can the MT system localize the errors, or must a human translator review both source and target texts in their entirety to determine the location of such errors? Unfortunately, errors committed by most MT systems span the gamut from innocuous to severe, and current systems seldom realize when they commit severe errors. Thus, a *95% FAHQT* system in the worst case produces a translated text that is analogous to a jar of cookies, only 5% of which are poisoned. Such a cookie jar is useless without a complete professional analysis to localize the poisoned ones.

The initial euphoria of the 1950's was followed by a grim realization that accurate translation requires some degree of text comprehension [2,1, 4]. As the MT problem proved to be much more complex than originally envisioned, the once lavish government funding programs were reduced to a trickle. At this point the MT community bifurcated into those who chose to address the fundamental problems of language understanding, helping to found the field of computational linguistics, and those who persevered in building MT systems. The latter group abandoned the unrealistic goal of developing fully automated translators and focused on the more pragmatic objective of building systems that increased the throughput efficiency of human translators. Several distinct approaches were taken; the most significant ones are discussed in the following section. More recently, newer technological developments are giving rise to qualitatively different methods. Section 3 discusses knowledge-based machine translation, the re-unification of the more theoretically motivated language processing methods with the objective of fully automated accurate translation. Section 4 outlines highly interactive, symbiotic human-computer approaches that promise to yield practical systems for low-volume, real-time translation.

Recent results indicate that the time may be finally coming to once again strive for the promise of true automated translation, fulfilling the aspirations of the early pioneers of the field.

2. Existing Approaches

Current machine translation systems range from translation aides that facilitate the job of a human translator to "best-effort" MT programs that require human intervention only after the fact -- in order to isolate and correct any errors committed in the automated translation phase. This section outlines the three major paradigms, assessing current and future potential.

2.1. Translation Aides

Much of the time of a human translator is wasted in manual lexicographic searches, and in document editing and formatting. Time consuming as they may be, these are the simplest tasks that a translator must perform, and therefore the easiest to automate effectively. Hence, one approach to improving the efficiency of a valuable, experienced human translator is to provide him or her with high-powered computational tools for the more mundane, time-consuming tasks. Such tools range from split-screen editing systems, to document formatters and graphic layout modules, to on-line technical dictionaries and grammar checking programs. Figure 2-1 outlines the basic flow of information in a machine-aided human translation approach.

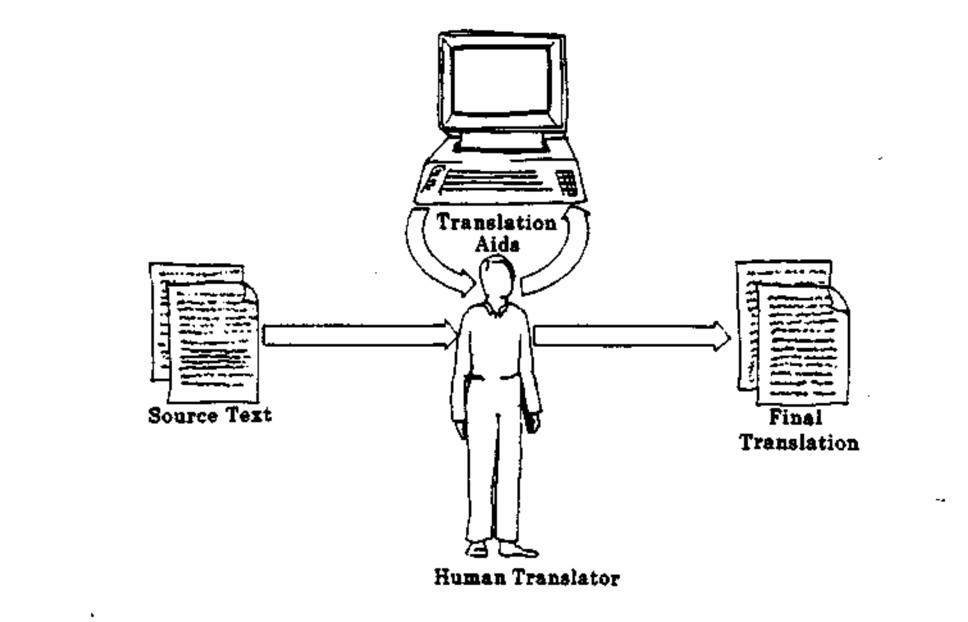


Figure 2.1: Translation Aids

Multiple versions of these tools are in existence [10], but much more productive sophisticated translation aids could be built, such as the following:

1. **Context-sensitive searching utilities** -- A human translator requires fast access to accepted technical terminology. To meet this need, on-line dictionaries have been implemented But, context-sensitive searching programs would make these far more effective. For instance, if a term has multiple meanings, these could be presented in rank order based on the topic of the text under translation, or based on previous terminological choices. Also, the ability to search for all occurrences of a technical term (or phrase) in the source text is very helpful. The translator may then decide to translate all occurrences identically, or to vary depending on local context.

2. **Automated dictionary update interfaces** -- No technical dictionary is ever complete, largely due to the rapid evolution of technical vocabularies vis a vis the slower evolution of general-purpose language. Thus, a dictionary needs to evolve with the language in an incremental manner. - The most effective way to track and stay abreast of a continuous but gradual lexical evolution is to enable the translators themselves to augment or modify dictionary entries. Such entries to local dictionaries can later be examined for eventual inclusion in the master dictionaries. However, experience has shown that building robust software to guide the translator in providing all the relevant information in the proper format is far from an easy task.[1]

[1] The ALEX facility of the LOGOS translation system is an example of a utility that provides a large fraction of the requisite functionality,

3 **Morphological analysis tools** -- The simplest aspect of automated language processing is morphological analysis (often coupled with secondary functionality such as spelling correction, etc.). Dictionary systems are far more effective when entries are stored in their basic form, and all inflections and other morphological variants are computed automatically. Of course, each language would require its own set of morphological analysis and composition rules, as well as exception tables.

The essence of all translation aids is that the human translator remains the central player, orchestrating all aspects of the translation process. The automated aids function only to increase efficiency (and possibly accuracy) by automating subsidiary tasks that would either be ignored or performed manually (such as searching several terminology banks for a possibly better translation of an obscure technical phrase, rather than manually searching several paper documents or simply accepting the translator's first guess). In contrast with the translation aids paradigm, all other approaches discussed in this paper place the automated system in the the central role, with the human checking results, correcting errors, preprocessing the input, or answering questions too difficult to be resolved automatically.

2. 2. Post-Editing Systems

Since fully-automated machine translation of unrestricted text has proven an elusive goal (as discussed earlier), several compromises have been made in automating as large a fraction as possible of the entire translation task. The most prevalent paradigm has been one of allowing an automated MT system do its best to translate unrestricted source text, and subsequently have a human translator (i.e., the *post-editor)* clean up the result. As illustrated in figure 2-2, systems requiring human post-editing of the translated output operate in the following manner:

1. The source text is converted to computer readable form.
2. The text is then sent to a batch-processing MT system, which produces a rough translation several hours (or days) later.
3. The original source text and the rough translation are presented to a human translator (i.e.. the post-editor), who cleans up the translation, fixing any errors or other difficulties.

Since post-editing requires significantly less time than complete translation, there is a potential for major gains in human efficiency. But, a knowledgeable human translator is still required. The post-editing approach has, until recently, predominated machine translation research and development. The domination had reached the the extent that adherents of the post-editing paradigm had on occasion considered all other approaches as temporary aberrations from the true path.

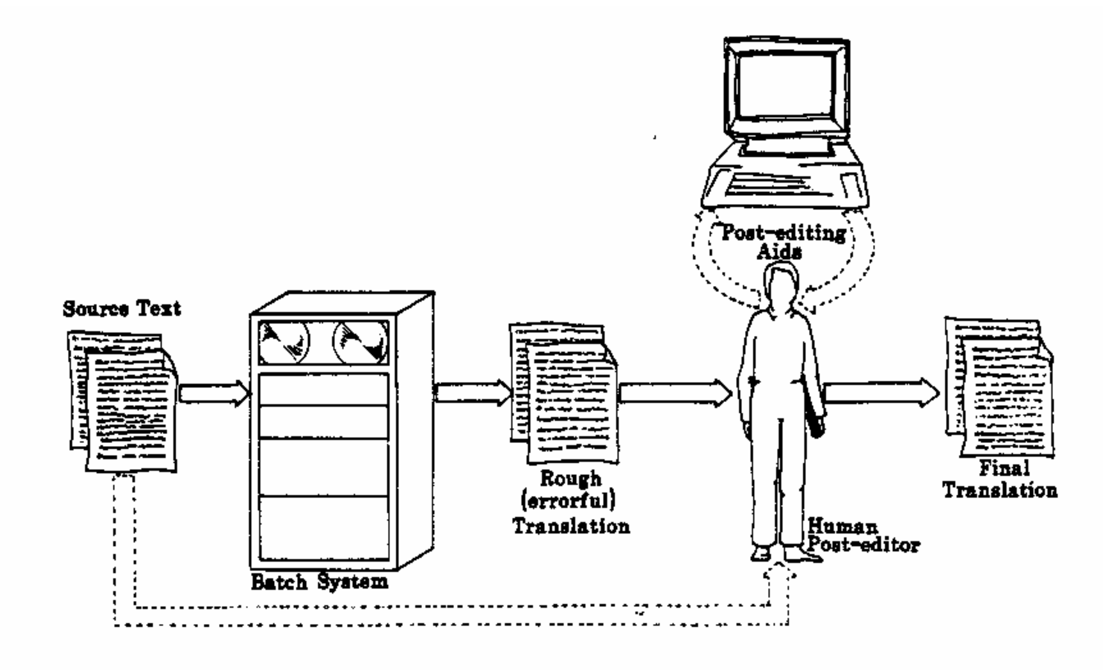


Figure 2-2: Post-editing Systems

2.3. Pre-editing Systems

The intervention of a human translator is required because current MT systems are unable to interpret all of the source text correctly. Errors in interpretation manifest themselves as incorrect translations. Thus, the post-editing approach recognizes the problem and attempts to minimize its impact by *post-facto* human corrections. An alternative is to ameliorate the problem at the interpretation phase by pre-editing the source text, eliminating difficulties such as complex grammatical structures, ambiguous words, and problematic semantic nuances. The pre-editing method is illustrated in figure 2-3.

The practicality of a machine translation systems is a function of accuracy and efficiency -- human efficiency being more significant than machine efficiency. Adherents of the post-editing approach have claimed that pre-editing is a time-consuming manual task, one that can also alter the meaning and intent of the source text in subtle ways. Thus, the general belief has been that pre-editing is less practical than post-editing on both counts: efficiency and accuracy. The veracity of this claim is difficult to ascertain in the general case, but for specific domains, such as translation of weather forecasts, pre-editing has proven quite viable [11]. The primary reason for the effectiveness of pre-editing in narrow domains is that source texts in such domains are usually written in their own jargon, in essence a fairly restricted sublanguage [12]. Owing to the relative simplicity of sublanguages, pre-editing can be held to a minimum, thus avoiding the problem of inefficiency and minimizing the

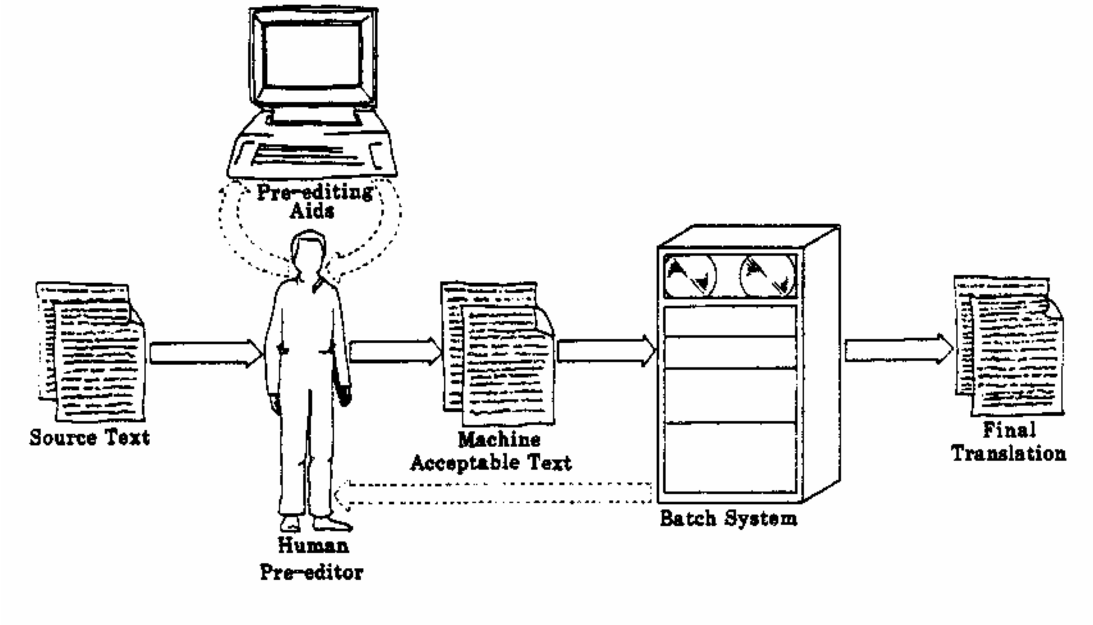


Figure *2-3:* Pre-editing Systems

problem of unwitting alteration of meaning. Nagao [14] suggested a Japanese sublanguage called “Machine Acceptable Language" where structural ambiguity is eliminated by extensive use of punctuation marks. In that system, pre-editing consists of manually inserting all the disambiguating punctuation into the source text.

3. Knowledge-Based Machine Translation

In order to address the *semantic invariance* criterion head on, a new approach to MT was developed, namely:

1. PARSE -- Map the source text into a language free meaning representation.
    a. Use a semantic knowledge base to disambiguate source text utterances, and to resolve other linguistic problems such as anaphoric referents.
    b. Encode only the meaning of the utterance, not its syntactic structure or source-text lexicon, in the semantic representation.
2. ELABORATE -- (Optionally) run a domain specific inferencer to fill in situational details left implicit in the source text
3. GENERATE -- Map the semantic representation into one or more target languages.

Knowledge-based machine translation (KBMT), depicted in figure 3-1, has been implemented in a

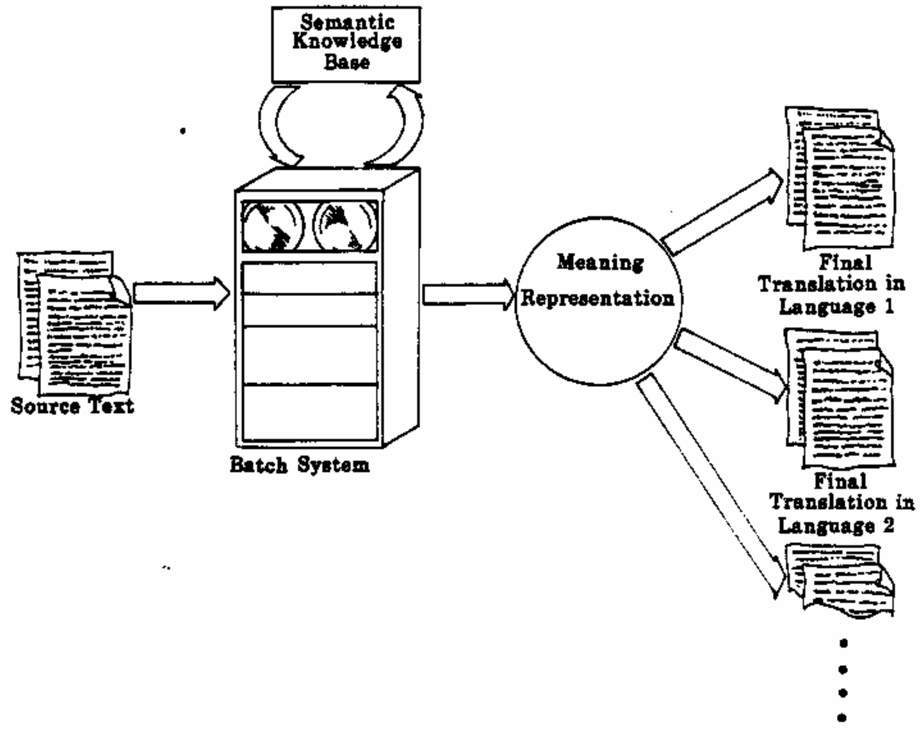


Figure 3-1: Knowledge Based Systems

pilot system called SAM[2] [4,8,17], and has proven successful in translating brief newspaper accounts of vehicular accident stories from English into Spanish, Russian, French, Mandarin Chinese and Dutch. SAM established the technical feasibility of KBMT -- as well as helping to reintegrate translation research as a mainline activity in the study of automated natural language processing. Newer and more robust semantic-based parsing techniques [9, 5] and better natural language generators [6] argue in favor of converting the KBMT approach from a laboratory exercise to production-quality translation systems in the very near future.

The semantic analysis required to build a language-free meaning representation has both advantages and drawbacks over earlier approaches. A clear advantage is that KBMT creates the possibility of true multi-lingual translation by the abandonment of *transfer grammars* in favor of more principled parsing and generation techniques. A transfer grammar [11,3] is a large, amorphous, *ad hoc* set of rules, referencing specific lexical entries, that map phrases in one language into corresponding phrases in another language. Thus, a complete transfer grammar needs to be created

[2] SAM, which stands for "Script Applying Mechanism", was a multi-faceted project originally conceived by Schank and Cullingford to explore the role of stereotypic domain knowledge on automated text understanding. Machine Translation occurred when several natural language generators were added to render its internal meaning representation into multiple languages.

for each *pair* of languages -- over 5,000 gargantuan grammars to translate between the 72 most active languages. The KBMT approach, however, requires only a parser to map the source language into the semantic representation and a generator to map that representation into the target language (72 parsers and 72 generators for any pair of languages in the example above). Moreover, if one text is to be translated into several languages, it need be parsed only once, and the resulting meaning representation generated in each target language. Generation is the simpler, less computationally demanding process. Thus, KBMT makes the process of multi-lingual translation far more computationally tractable -- as well as reducing significantly the amount of development work required to reach eventual closure in the number of grammars needed to translate among all commonly-spoken human languages.

Perhaps the major disadvantage is that the KBMT process produces a paraphrase of the source text in the target language, rather than performing "exact" translation -- in the sense that it does not strive to achieve lexical or syntactic invariance. Thus, knowledge-based machine translation would be singularly inappropriate for translating poetry or other literary forms where the very structure of the text conveys a central message. Moreover, KBMT requires general semantic information and domain-specific knowledge roughly proportional to the semantic knowledge base that a human translator would bring to bear. With this caveat, KBMT could become highly practical for domains where *a* large volume of material must be translated swiftly and accurately, but less practical for low volume domains where it is more difficult to amortize the cost of building the domain-specific knowledge bases.

Finally, we should stress that if a meaning representation can be constructed automatically and unambiguously, the poisoned-cookie problem does not arise. Unlike the older post-editing approach, no human translator is needed to read carefully both source and target texts to determine where the meaning was radically altered. If KBMT can translate at all, meaning remains invariant.

4. The Interactive Approach

The interactive approach illustrated in figure 4-1 is particularly suitable in systems in which an input text is provided directly by the user. In this approach, the user types a sentence (or a text) in his language; the system asks him questions in his language whenever needed; the user answers those questions; and finally the system produces a sentence (or a text) in the target language which does not require post-editing.

The interactive approach is especially desirable when:

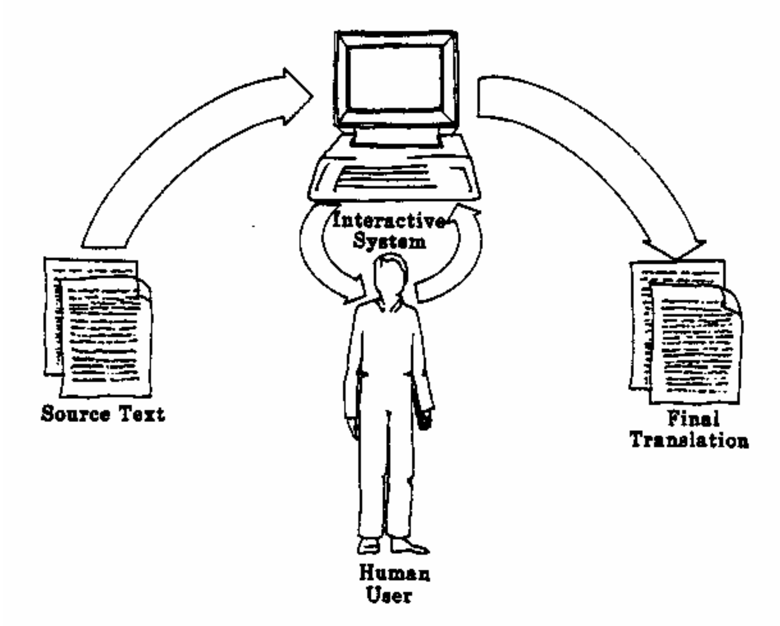


Figure 4-1: Interactive Systems

- A text is so small and personal that the user cannot afford calling a translation service, or the urgency of the task precludes waiting days for the result.
- A text is to be translated into several languages (as with the KBMT approach).
- A text is so technical that even professional translators require help from domain experts.

The first situation is discussed in detail by Tomita [20]. The second and third situations were advocated by Kay [10].

Unlike conventional post-editing systems, the interactive approach exhibits the following characteristics.

- The user does not have to know the target language.
- The user need not have any special knowledge of linguistics, software, translation, etc.
- The system's final output requires no post-editing.
- Thus, everybody can use the system and generate target language texts without assistance of a translator or post-editor.

In designing and implementing an interactive translation system, the following characteristics are highly desirable

- Because the system must run in real time rather than as a batch job, its response time should be reasonably quick.

- Because the system's input can be typed in from the terminal and not provided as a polished text file, the system must be reasonably robust against ill-formed sentences.

- The system, being particularly well suited for small, rapid turn-around, but possibly infrequent tasks, must run on affordable general purpose machines (such as high-end micros), so that it can be used at home or office on demand.

4.1. Interactive Sentence Disambiguation

This subsection describes how to resolve sentence ambiguity by asking the user focused questions. The essence of the interactive approach is to bypass the massive semantic knowledge requirements of KBMT by querying the user to disambiguate troublesome sentences. Such disambiguation, however, should not presuppose any formal linguistic, computer science or target language knowledge on the part of the user.

To resolve word-sense ambiguity, a system can ask questions such as the following:

```
The word "pen" means:
1) a writing pen
2) a play pen
NUMBER?>
```

To resolve referential ambiguity, a system can ask in the following manner:

```
The word "she" refers:
1} "Cathy"
2) "my mother"
3) "the sailboat"
NUMBER?>
```

Those two kinds of interactive disambiguation can be implemented relatively easily by simply enumerating alternatives on the screen. However, resolving syntactic ambiguity is not that easy. It is clearly not acceptable to simply enumerate all possible parse trees on the screen, because:

- the user may not be familiar with tree structures, and

- the number of alternatives can number in the hundreds [7].

Therefore, we need a little more intelligent mechanism as in the following example:

```
I saw a man and a woman with a telescope.
1) "a man" and "a woman"
2) "a man" and "a woman with a telescope"
NUMBER?>


1) the action "saw" takes place "with a telescope"
2) "a man and a woman" are "with a telescope"
NUMBER?>
```

The algorithm for this interactive disambiguation was first introduced in [18], and a polished version is described recently in [21].

Tomita et al. [19] built an experimental interactive system, modifying Nishida and Doshita's English-Japanese machine translation system [15] so that the system is capable of asking questions interactively to disambiguate its input sentences. Experiments show that in general, the syntactic ambiguity of a sentence can be resolved by a couple of questions, assuming that a little semantic knowledge is available (so that the system can resolve the simpler ambiguities by itself) [20].

4.2. Bypassing Source Text

This subsection describes a different kind of interactive systems that generate a target language text by interactive dialog with the user, requiring no source text. Of course, these systems are more "automated text composition systems" than true machine translation systems. In such systems, questions are asked to construct the "semantic content" which contains enough information to generate the target text, as illustrated in figure 4-2.

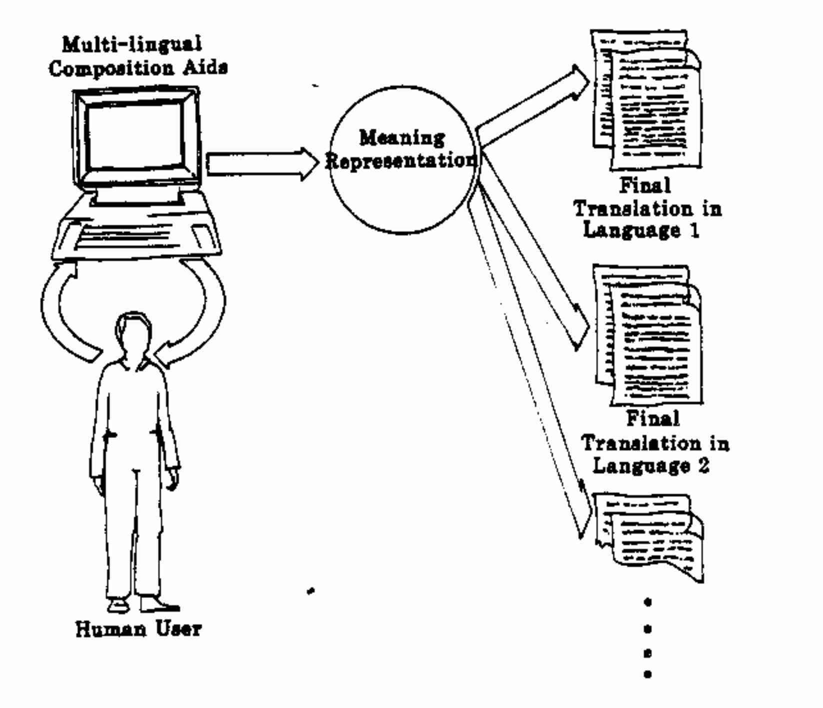

Multi-lingual Composition Aids
Meaning Representation
Final Translation in Language 1
Final Translation in Language 2
Human User

Figure 4-2: Multi-lingual Composition Aids

This paradigm can be thought as one extreme variation of the interactive method in which the system obtains semantic content directly from the user without parsing any source text. Saito and Tomita [16] built a prototype system that enables the user to generate formal letters in several languages only by interacting with the user in his language. Although the system can handle only stereotypic topics such as moving notification, the quality of its output is so good that the user might want to use the system to produce letters in his own language as well.

First the user is asked his language and the language of the target letter:

```
  Your  language ?
1   English
2   Japanese
3   Spanish
4   French
5   German
1 - 5 ?  1

  Target language ?
1    English
2    Japanese
1- 2 ?       2
```

Next the user is asked the topic of a letter he is writing:

```
  The topic of a letter ?
1   Moving
2   Thanks for Gift
3   Invitation
4   Happy New Year
1   -  4  ?     1
```

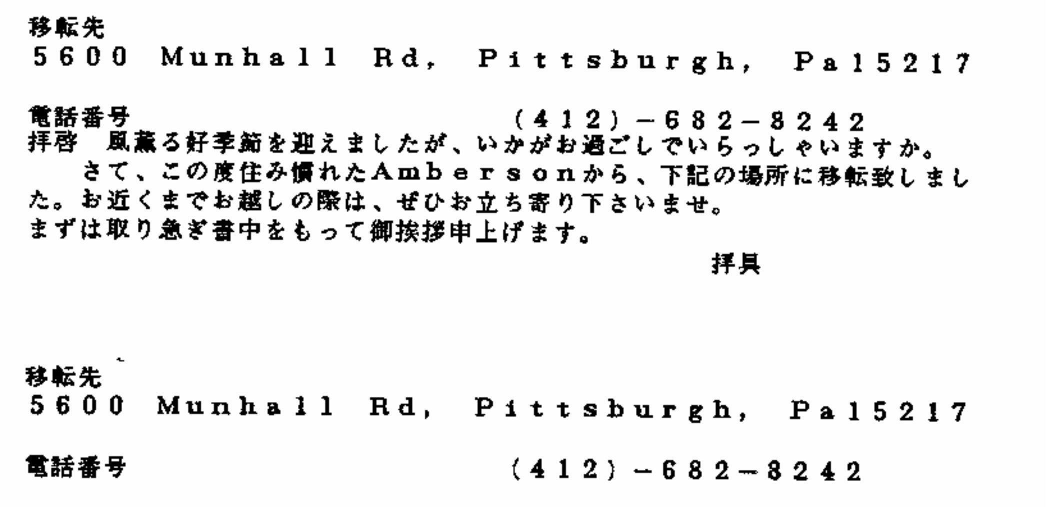
The system then asks the user questions to acquire enough information to generate a moving notification letter in Japanese.

```
  What's your old address ?
Type --->       Amberson

  What's your new address ?
Type  --->        5600 Munhall Rd, Pittsburgh, Pa 15217

  What's your new phone  number ?
Type  --->        (412)-682-8242

  To whom are you writing ?
1   Business acquaintance
2   Superior
3   Friend
1 - 3 ?  2

  What month is it now ?
1 - 12 ?  6

  Have you finished moving ?
Y / N ?  y
```

The completed semantic content is illustrated below in a simple frame representation:

```
[ Moving
    [ writetowhom: superior                                  ]
    [ nowseason: june                                        ]
    [ fromwhere: Amberson                                    ]
    [ towhere: 5600 Munhall        Rd, Pittsburgh, Pa15217   ]
    [ tel: (412)-682-8242                                    ]
    [ done: yes                                              ]
]
```

The final and Japanese text generated by the generator out of the semantic content is shown below.



移転先
５６００　Ｍｕｎｈａｌｌ　Ｒｄ，　Ｐｉｔｔｓｂｕｒｇｈ，　Ｐａ１５２１７

電話番号　　　　　　　　　　　（４１２）－６８２－８２４２
拝啓　風薫る好季節を迎えましたが、いかがお過ごしでいらっしゃいますか。
　　さて、この度住み慣れたＡｍｂｅｒｓｏｎから、下記の場所に移転致しました。お近くまでお越しの際は、ぜひお立ち寄り下さいませ。
まずは取り急ぎ書中をもって御挨拶申上げます。
　　　　　　　　　　　　　　　　　　　　　　　　　　拝具

移転先
５６００　Ｍｕｎｈａｌｌ　Ｒｄ，　Ｐｉｔｔｓｂｕｒｇｈ，　Ｐａ１５２１７

電話番号　　　　　　　　　　　（４１２）－６８２－８２４２

5. Concluding Remarks

Whereas this paper has focused on well-defined paradigms for machine translation, we do not mean to rule out hybrid approaches. In fact, the combination of two or more approaches may prove superior in many circumstances. For instance, the knowledge-based and interactive approaches may be combined as illustrated in figure 5-1. For most routine semantic decisions, the combined system queries its knowledge base. On the rare occasions when that query proves insufficient (e.g., the topic of the text strayed from its expected domain to one where the system lacks knowledge, or the system's knowledge is otherwise incomplete), the interactive component formulates a focused question to the user. Such compromises may prove to be the key to practicality, if neither extreme proves feasible.

Having surveyed the major approaches to machine translation, we observe that the established post-editing technique has received the most commercial attention, despite some of its more obvious weaknesses. Some of the newer approaches, such as KBMT, are based on recent developments in computational linguistics and artificial intelligence, such as semantic analysis and knowledge representation techniques. Thus, they have not yet emerged from the laboratory to be tested in a

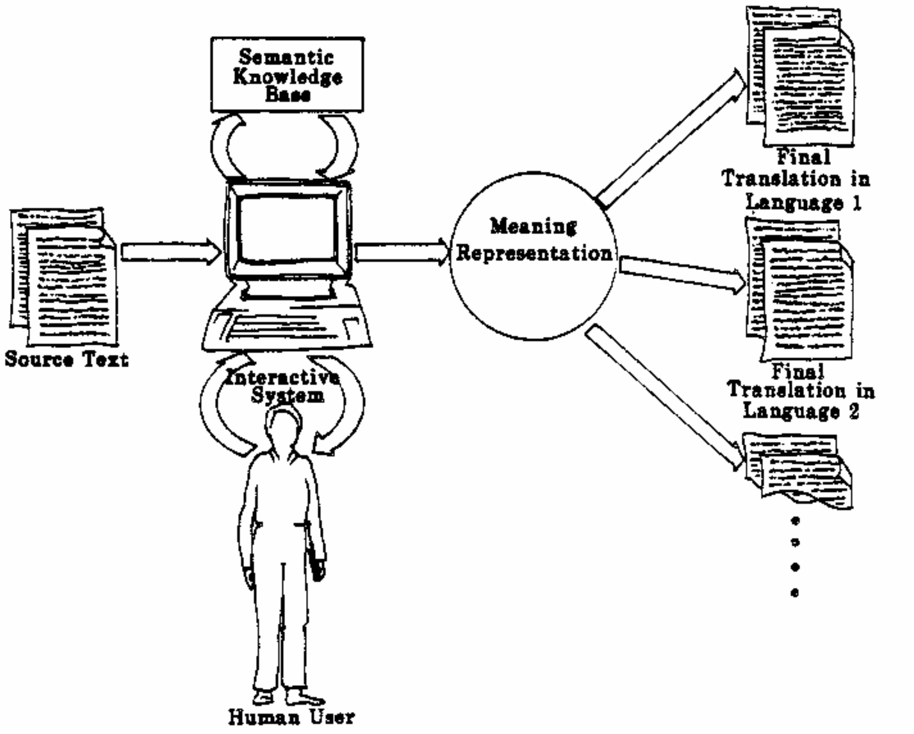


Figure 5- 1: Knowledge Based Interactive Systems

: eduction environment The interactive approach requires sophisticated interactive computers for practical application; such machines are just recently becoming widely available. Hence, in the near future we should be able to produce practical systems based on these newer, more powerful techniques. Perhaps by that time still newer methods may be brewing in our research centers, based on better understanding of linguistics, knowledge representation, and computational techniques.

6. References

1. ALP AC, *Language and Machines,* National Academy Science, 1966.
2. Bar-Hillel, Y., "The Present Status of Automatic Translation of Languages," *Advances in Computers,* Vol. 1, 1960, pp. 91-163.
3. Boitet, C., "Problemes actuels en TA: Un essai de response," Proc. *6th International Conference on Computational Linguistics,* Ottawa, Canada, 1976.
4. Carbonell, J. G., Cullingford, R, E. and Gershman A. G., "Steps Towards Knowledge-Based Machine Translation," *IEEE Trans. PAMI,* Vol. PAMI-3, No. 4, July 1981.
5. Carbonell, J. G. and Hayes, P. J., "Recovery Strategies for Parsing Extragrammatical Language," *American Journal of Computational Linguistics,* Vol. 9, No. 3-4, 1983, pp. 123-146.

6. Carbonell, J. G., Boggs, W. M., Mauldin. M. L. and Anick, P. G., "The XCALIBUR Project, A Natural Language Interface to Expert Systems and Data Bases," in *Applications in Artificial Intelligence, S.* Andriole, ed., Petrocelli Books Inc., 1985.

7. Church, K. and Patil, R., "Coping with Syntactic Ambiguity or How to Put the Block in the Box on the Table," Tech. report MIT/LCS/TM-216, Lab. for Computer Science, Massachusetts Institute of Technology, April 1982.

8. Cullingford, R., *Script Application: Computer Understanding of Newspaper Stories,* PhD dissertation, Yale University, Sept. 1977.

9. Hayes, P. J. and Carbonell, J. G., "Multi-Strategy Parsing and its Role in Robust Man-Machine Communication," Tech. report CMU-CS-81-118, Carnegie-Mellon University, Computer Science Department, May 1981.

10. Kay, M., "Machine Translation," *American Journal of Computational Linguistics,* Vol. vol.8, No. no.2, April-June 1982, pp. 74-78.

11. Kittredge, R., Bourbeau, L. and Isabelle, "Design and Implementation of an English-French Transfer Grammar," *Proceedings of the 6th International Conference on Computational Linguistics,* Ottawa, Canada, 1976.

12. Kittredge, R. and Lehrberger, J., *Sublanguages: Studies of Language in Restricted Semantic Domains,* deGruyter, Berlin, 1981.

13. Locke, W. N. and Booth, A. D., *Machine Translation of Languages,* Technology Press, Boston, MA, 1957.

14. Nagao, M., "On Restricted Sublanguage," *IPSJ Symposium on Natural Language Processing,* 1983.

15. Nishida, T, and Doshita, S., "Application of Montague Grammar to English-Japanese Machine Translation," *Proceedings of Conference on Applied Natural Language Processing,* 1983, pp. 156-165.

16. Saito, H. and Tomita, M., "On Automatic Composition of Stereotypic Documents in Foreign Languages," Tech. report, Computer Science Department, Carnegie-Mellon University, 1985.

17. Schank, R. C. and Abelson, R. P., *Scripts, Goals, Plans and Understanding,* Hillside, NJ: Lawrence Erlbaum, 1977.

18. Tomita, M., "Disambiguating Grammatically Ambiguous Sentences by Asking," *Proceedings of 10th International Conference on Computational Linguistics (COLING84),* 1984.

19. Tomita, M., Nishida, T. and Doshita, S., "User Front-End for disambiguation in Interactive Machine translation System," *IPSJ Symposium on Natural Language Processing (in Japanese),* 1984.

20. Tomita, M., "Feasibility Study of Personal/Interactive Machine Translation Systems," *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* Colgate University, August 1985.

21. Tomita, M., *An Efficient Context-free Parsing Algorithm for Natural Languages and Its Applications,* PhD dissertation, Computer Science Department, Carnegie-Mellon University, May 1985.