English - Malay Translation System : A Laboratory Prototype

TONG Loong-Cheong

Computer Aided Translation Project
School of Mathematical and Computer Sciences
Universiti Sains Malaysia
11800 Penang, MALAYSIA

Abstract

This paper presents the results obtained by an English to Malay computer translation system at the level of a laboratory prototype. The translation output obtained for a selected text (secondary school Chemistry textbook) is evaluated using a grading scheme based on ease of post-editing. The effect of a change in area and typology of text is investigated by comparing with the translation output obtained for a University level Computer Science text. An analysis of the problems which give rise to incorrect translations is discussed. This paper also provides statistical information on the English to Malay translation system and concludes with an outline of further work being carried out on this system with the aim of attaining an industrial prototype.

1. The English to Malay Translation System

Background

Computer Aided Translation (CAT) research at Universiti Sains Malaysia (USM) began in 1976 as an individual research effort. However, at that time, the work is more appropriately classified under natural language data processing, including topics such as 'istilah' (terminology) information retrieval, Malay rootform extraction, parsing of Malay sentences using context-free grammars and Malay language teaching tools [Tong 78, Chang 78].

In 1978, research into CAT was initiated, and by 1979, the researchers at USM began to develop grammar models for English to Malay translation using the software tool ARIANE [GETA 78].

In 1980, a national workshop was conducted in USM, where a pilot English to Malay translation system was demonstrated. Financial support became available, and further development on the basic translation model was carried out [Tong 82, van Klinken 84, Zaharin 84].

In 1984, a permanent Computer-Aided-Translation Project unit was set up at USM, and full-time research staff were assigned to this project. Members of this project group now include two computer scientists, one linguist, and five lexicographer / editor / terminologist. This group was assigned the task of producing a laboratory prototype for English to Malay translation, and the result of their efforts is presented in this report.

System Environment

The ARIANE system is an integrated software environment for computer-aided-translation, including tools for compiling grammars and dictionaries, and for processing corpus of the source and target texts. The CAT concepts behind this system is well-known and well-documented [Boitet and Vauquois 1985].

This software has been programmed using different levels of computer languages, from IBM assembly (PL360) to PL/I, and making extensive use of system tools of the IBM VM/CMS system - XEDIT and EXEC. One of its advantages is efficiency (as compared to other similar systems), which means that it can execute with reasonable speed even on a comparatively small computer system. USM's experience with the ARIANE system has been very satisfactory, and we doubt very much if another system could have been migrated and utilised at this University with similar success. Although there had been some criticisms about ARIANE in the literature, our experience has shown that inspite of its recognised weaknesses and drawbacks, it remains an extremely powerful and practical set of tools for the development of CAT systems. Of course, the methodology pioneered at GETA [Vauquois 75] has been incorporated into many 'new' systems today.

On the physical side, the ARIANE system itself occupies about 8 Mbyte of secondary storage, while the user machine requires another 5 Mbyte for storing the linguistic data (grammar models and dictionaries, but not including the source and target texts and their intermediate results). A virtual memory size of 2 Mbytes is used for the execution of all the translations from English to Malay described in this report.

Translation Model and Execution Time

The English to Malay translation system consists of three main dictionaries - source English, English-Malay transfer, target Malay - and five grammar models. The size of these various components are as follows:

Dictionaries:

Source lexicals:	5,000	(≃11,000 words)
Target lexicals:	4,000	(≃ 9,000 words)

Grammar models:

	lines	rules
morphological analysis	600	90
structural analysis	5600	300
structural transfer	800	47
structural generation	1700	120
morphological generation	900	120

The execution time for translation is estimated at 1.0097 MIPW (million of instructions per word). This is consistent with times measured at GETA, Grenoble [Boitet and Vauquois 84]. In practical terms, this means that on USM's IBM 4381 system (estimated at 2.1 MIPS), the translation time is approximately 0.48 second of virtual CPU time per word. This figure is based on the translation time for about 3,000 words taken from the selected text. The proportionate time for each phase of the translation process is as follows:

	percent
morphological analysis	0.33
structural analysis	55.21
lexical transfer	0.44
structural transfer	11.34
structural generation	31.47
morphological generation	1.21

From the above, it can be seen that the three dictionary retrieval phases together account for only 2 % of the time, while the structural analysis phase used up more than half the total time, with the rest taken up by the structural generation (about one-third) and the structural transfer phases. This result is

again consistent with those for other translation models at GETA, Grenoble.

2. The Quality of Translation

Grading Scheme

In order to assess the 'quality' of the translation output, a grading scheme (from grade A to grade F) was devised using a sentence as the boundary of assessment. This scheme is based on the ease of post-editing the translation output, and not on the quality or standard of translation in the more usual sense.

Currently, there is no established method of evaluating computer-aided-translation or mechanical translation output. Ease of post-editing is a measure which also takes into account the ease of understanding as well as the accuracy of translation. Two important factors which affect any grading scheme is the typology of the source text itself and the expert knowledge of the evaluator in that particular area of text. Some method of evaluating the ease of understanding of the source text and some definition of a neutral evaluator are prerequisites to any standardised evaluation scheme.

The grading scheme proposed in this report is a measure of the time required to edit sentences translated by the computer, ranging from fast (as in grade A where no post-editing is required) to slow (as in grade F where a sentence has to be retranslated manually). There has been no attempt to categorise the source sentences into different degrees of difficulty or length. Hence, the typology of text used in this evaluation must be borne in mind when assessing the overall results. Although grades are assigned to individual sentences, the source texts were extracted by paragraphs, and hence, the continuity of the text is maintained. The actual grading itself was carried out by more than one individual in order to reduce (as much as possible) the effect of individual 'bias'. After careful scrutiny, it was concluded that variation in the results obtained is within expected limits, thus allowing broad conclusions to be drawn on the effectiveness/usefulness of the translation system.

The grades assigned to translated sentences are as follows:

- A: correct translation, no modification required.
- B: list of alternative words selected by post-editor.
- C: understandable translation (with preservation of meaning), single word corrections without reference to source text.
- D: as in C, but reference to source text is necessary.
- E: major modifications with reference to source text.
- F: retranslated manually.

Results for Selected Area and Text Typology

A Chemistry textbook for upper secondary school was chosen as the first text for the development of the laboratory prototype. A total of 393 sentences were extracted at random from this textbook and translated by the computer. The translation output is then graded by three human post-editors and the result given below is based on their combined evaluation.

Grade:	A	B	C	D	E	F
No. of sentences	61	125	114	85	8	0
Percentage %	15	32	29	22	2	0
Cumulative %	15	47	76	98	100	100

The above result shows that 76 % of translated sentences are 'understandable' (no reference to English source text is necessary) and requires, at the most, only minor modifications during post-editing.

Effect of a Change in Area and Text Typology

The new text is a University level Computer Science textbook, from which 207 sentences were extracted, translated by the computer, and then graded. The result is as follows:

Grade:	A	B	C	D	E	F
No. of sentences	23	44	74	41	11	14
Percentage %	11	21	36	20	5	7
Cumulative %	11	32	68	88	93	100

As expected, the quality of translation in this case is lower than that for the Chemistry text. Most of the additional problems encountered can be solved either through dictionary coding or minor modifications in the grammar. With these changes, the quality of translation for the Computer Science text is expected to be raised to the same level as that for the Chemistry text.

3. Existing Problems Classification

An attempt was made to analyse the problems encountered, i.e. the errors in translation output. This involves a tedious process of correctly identifying the source of each error found in the translation output, and then classifying them according to the phase of translation (i.e. analysis, transfer or generation) at which they occur. The purpose is to identify simple problems which can be solved in the existing system through modifications to the linguistic data, while more complex problems can be the subject of further research. This analysis of errors also provides statistical information on their distribution and importance, hence giving some guidelines as to their priority for further investigation.

The Analysis Phase

The problems of ambiguity and coordination account for more than half of the errors at the analysis phase. The problem of ambiguity here refers to ambiguities which remain unresolved at the end of analysis and to cases of erroneous disambiguation. This type of problem is by far the most important, accounting for close to 50 percent of the existing errors found in the analysis phase.

Ambiguities which remain unresolved include

verb/noun	('form','works','use'),
verb/adjective	('direct','total'),
verb/ven	('.. is unglazed paper..'),
noun/adjective	('routine','plural'),
verb/ving	('.. painting of...'),
adj/pronoun	('other').
'as'	

Coordination (apposition, inclusion) is a serious structural problem not handled particularly well by the existing grammar model. Many different types of elements can participate in coordination (apposition, inclusion) and examples of cases not considered in the current grammar are:

complex noun phrases,	
prepositions,	('to and from and within..')
verbal clauses,	('...but.....and.....')
interrogatives,	('why....and do....')
adjunct phrases.	('..hot and humid..')

Other errors in the analysis phase are relatively less complex and can be solved through modifications or improvements in the morphological and structural analysis grammars and in the coding of the source dictionary. Errors in this category are:

- errors in morphological coding, including idiomatic expressions and compound words;

- 'unknown' structures in the current model, such as

(elision)
'., although large enough to pass through..'

(embedded imperative)
'; hence the instruction: shake the bottle.'

(complex comparative)
'..the same temperature as that at which..',

(enumeration)
'.... only 4 operations:
I/O, arithmetic, comparison, movement of data.'

Various bugs still exist in the analysis grammar model itself and these will be corrected as part of the maintenance on the translation system.

The Transfer Phase

The two main problems at the transfer phase are the incomplete (or incorrect) choice of target lexicals, and the transfer of idiomatic expressions.

The disambiguation of a s ource lexical which carry more than one meaning and which is translated by different target lexicals accounts for more than half of the errors at transfer. The source of this problem is actually at the analysis phase, which was unable to produce a sufficiently deep level of interpretation (e.g. semantics and semantic relations) to solve the ambiguity which manifests itself only at transfer.

The two categories of words which are most problematic are the verbal forms ('reveal', 'assume', 'call') and the prepositions ('in','by','to'). Although disambiguation rules based on context are employed during the structural transfer phase, they can only solve relatively straightforward cases. For the more difficult cases, the current approach of displaying a list of multiple choices of words to the human post-editor seems to be the most acceptable solution. Much deeper work in state semantics and semantic relations will have to be carried out in order to improve on this. Even if such improvements are found, there is still the question of weighing the cost of such sophisticated processing by the computer (which is expected to be very high) against the cost of human post-editing.

Idiomatic expressions are normally coded directly in the source dictionary. Unfortunately, the ARIANE software does not provide sufficient facilities at analysis or at transfer phase to cater for some of the complex manipulations required. Some idiomatic expressions are ambiguous (i.e. they can be considered idiomatic only in certain context), and hence, there is the problem of disambiguating them during analysis. Also, some English idiomatic expressions are particularly difficult to translate into Malay, and perhaps other target languages as well.

The Generation Phase

Errors during structural generation are relatively few, and also relatively minor from the point of view of post-editing. Most errors during this phase will give rise to grade C sentences if there are no other type of errors in the sentence.

The main problems are as follows:

** Position of elements in complex noun phrase.

Most of the errors are due to the incorrect placement of the preposition 'bagi' (similar to 'of' but not as commonly used) in a complex Malay noun phrase. Other elements of the noun phrase which give rise to errors are the '-ing' or '-en' form used as an adjective, and the lexicals 'other' and 'only' which seem difficult to translate into Malay. Very often, an adjective in Malay is introduced by the relative pronoun 'yang'. However, there seems to be no consistent rule for this. Certain lexicals always require a 'yang', while others only under certain not well-defined conditions.

** Position of adverbs and adjuncts of clauses.

This problem is not very well investigated in the existing model, and can hopefully be improved upon at a later stage.

** Relative clause introduced by a preposition.

The relative clause introduced by a preposition ('in which', from whom', etc.) is particularly difficult to translate into Malay (even for human translator). Formal linguistic study is being carried out into possible target structures. This is one specific case whereby linguistic research is initiated specifically to cater for the needs of computer-aided-translation.

** The generation of Malay pronouns.

Another difficult problem is the translation of some pronouns - 'it', 'they', 'another', 'one', 'latter', 'former ', 'those'. The Malay language sometimes requires a repetition of the referenced object in place of the pronoun. Even when this is not necessary, as in the case of a pronoun referring to an undefined object, it may be incorrect to translate directly with the equivalent pronoun ('ia', 'mereka', 'yang lain', 'kita'). Again further investigation into the linguistic aspects of this problem will be necessary before an acceptable solution can be found.

source: 'move from one part of the solid to another'

computer: 'bergerak dari 1 bahagian pepejal kepada yang lain'

edited: 'bergerak dari 1 bahagian pepejal kepada bahagian yang lain'

4. Further Work on the Laboratory Prototype

Grammar Model Development

Many problems remain to be tackled both from the linguistic as well as the computer science point of view. Some of these problems, especially at the generation phase, are at the surface or syntactical level. Further work on the grammar model should bring about improvements in this area.

The problems of coordination during analysis and lexical ambiguity during transfer are at a deeper semantics level. Until formal linguistic work on semantics (such as Montague Grammar) can come up with some practical solution, these problems are only amenable to a linguistic engineering approach based on some static categorisation of semantics together with some generalised dynamic method of processing and the ability to handle exceptional cases.

The current English analysis model already contains a very

comprehensive set of disambiguation rules. For the more difficult cases which still remain unresolved at the end of analysis, exhaustive search method can be employed. This is not as costly as it may seem, since a survey of such cases has indicated that good heuristic conditions are possible to reduce the overall search time.

The current analysis model attempts to achieve a deep level of interpretation right up to logical and semantic relations. Since this level may not be attainable for many of the sentences in a particular text, a lower level of interpretation such as syntactic functions or even morphosyntactic classes should be used instead. A large proportion of such sentences can still be translated correctly, and therefore, the provision of this 'safety net' is essential.

Dictionary Coding

The development of an industrial prototype will demand a considerable increase in the size of the dictionary, at least to about 10,000 source lexical units. Hence, lexicographic work represents the single most important and time-consuming task in the development of an industrial prototype. Preparations are already underway to simplify this task by producing a simplified form (or questionnaire) which can be filled-up by lexicographers with perhaps only a minimal amount of training. Data from this 'form' can then be transferred into computer codes to be used by the translation system.

This preparation of a computerised dictionary can also be integrated with any work being carried out on lexical databases for ordinary human consumption. The two tasks have a large amount of intersecting information needs, and hence, can be mutually beneficial.

Towards an Industrial Prototype

The laboratory prototype is now ready for development into an industrial prototype. The first task is of course the drawing up of a list of possible applications, followed by a feasibility study of the text typology for each of these application. The final selection will be based on the quality of translation which can be expected and the type of financial support available. Other important considerations include:

- the volume of translation work,
- the frequency of translation work,
- the urgency / speed of the translation work,
- the availability of a complete set of Malay technical terms,
- the availability of text materials in machine-readable format.

Once an application has been selected, the next step is the organisation of the development work itself. Here, the available manpower is a critical element, and from experience, it is very difficult to convince policy makers and financial supporters on this. Any development team must be made up of high-calibre computational linguists, computer scientists, lexicographers, editors and translators, who must be well-trained in the methodology of computer-aided-translation besides their own area of specialisation.

Another important factor for planning purposes is the time required to develop an industrial prototype, and this has also been frequently underestimated. It is estimated that at least 3 years work by the existing research team at Universiti Sains Malaysia will be required to complete an industrial prototype for English - Malay translation in one specific area of application.

A Dedication

Without the late Professor B. Vauquois, the CAT project at Universiti Sains Malaysia would not have existed. His dedication has inspired all who worked with him, and his kindness will always be remembered.

References

1. [Boitet and Vauquois 84]
Christian Boitet and Bernard Vauquois
'Automated Translation at GETA'
GETA, Aug 1984.

2. [Chang 78]
Chang May See
'Computer System Aids in Natural Language Data Processing'
M.Sc. Thesis, USM, Oct. 1978.

3. [GETA 78]
M. Quezel-Ambrunaz
'ARIANE 78: Systeme interactif pour la traduction automatique multilingue'
Tech. Report GETA, Sep 1978.

4. [Tong 78]
Tong Loong Cheong
'An Information Retrieval System with Linguistic Capability'
Proc SEARCC, Sep 1978.

5. [Tong 82]
Tong Loong Cheong
'Computer Aided Translation - Technical Report Compilation'
Tech. Report PTMK, Dec 1982.

6. [van Klinken 84]
Catharina van Klinken
'Disambiguation Strategy in English Structural Analysis'
Tech. Report PTMK, Dec 1984.

7. [Vauquois 75]
Bernard Vauquois
'La traduction automatique a Grenoble'
Documents de linguistique quantitative,
DUNOD, 1975.

8. [Zaharin 84]
Zaharin Yusof
'The Morphological Generation of Malay'
Tech. Report GETA, Oct. 1984.