[*Machine Translation Summit*, September 17-19, 1987, Hakone Prince Hotel, Japan]

The Current Stage of The Mu-Project

TSUJII, Jun-ichi
Dept. of Electrical Engineering, Kyoto Univ., Kyoto, Japan

1 Introduction

The Japanese government MT project (the MU project), which was a four year project supported by the special coordination fund of STA (Science and Technology Agency), was completed at the end of March, 1986. The main objective of the project was to develop a proto-type system which would demonstrate the feasibility of MT systems for certain restricted subject domains and document types. We considered that this initial objective of the project was well fulfilled so that we started another four year project (the MU-II project) from April, 1986. The new project aims to develop & system which will be used for daily translation services in JICST (Japan Information Center for Science and Technology) from 1990.

Because the technical details of the MU-project have been given in the other papers (1) (2) (3) (4) (5), we will report overall results of the Mu Project and brief introduction of the outline of the new project.

2 Outline of the Mu Systems

The following gives the rough sketch of the current MU systems.

1. Basic Approach: Transfer Approach
2. Design Principle: lexicon Driven Processing
   Neutral Dictionary
   Heuristically Guided Processing
3. Language Pairs: Japanese to English, English to Japanese
   Two systems have been developed, both of which uses same software systems
4. Subject Fields: Electrical Engineering
   The subject fields will be extended in the other scientific and technological fields in the new project.
5. Document Type: Abstracts of Scientific and Technological Papers
6. Implementation Language:
   - Core translation system:
     - Uti-LISP (University of Tokyo Interactive Lisp) with some augmentations for treating Chinese characters, etc.

- Zeta-LISP (Symbolics 3600 series) The main systems were developed first by Uti-LISP on FACOM 382, and then, they were converted into the Zeta-LISP version.

• Support Systems including interactive editing systems for texts and dictionaries, management of the dictionary data bases, etc: PLI

7. Grammar Writing Language: A special language called GRADE was designed and implemented by LISP for writing grammars in the project. GRADE facilitates flexible tree structure matching, backtracking, etc.

8. Size of Dictionaries: Verbs, Adjective, Adverbs, etc. = 5,000
Nouns (incl. terminology) = 70,000

3 The Transition Process and the Size of the Grammars:

We just show the size of the English-Japanese system. The system of the other direction is of almost same size.

3.1 English Analysis Grammar:

- Total Number of SG = 500
- Total Number of RR = 2298

The following are the main steps of analysis which are performed in the order shown.

Step 1 Post-Morphological Analysis including the processing of number expressions, parenthetical expressions, compound words, etc. (SG = 52, RR = 138)

Step 2 Treatment of Question and Imperative sentences (SG = 8, RR = 20)

Step 3 Disambiguation of parts of speech based on local environment (SG = 68, RR = 374)

The decision of parts of speech at this step is tentative. The rules in this step are heuristic ones only based on local environments around the ambiguous words, which are expected to give correct interpretations for most cases.

However, because the decisions given at this stage are sometimes wrong, the following stages revise the decisions (see below).

Step 4 Adverbial Phrases (SG = 4, RR = 43)

Step 5 Simple Noun Phrases (Phrases whose heads are nouns and which consist of prenominal modifiers and the heads) (SG = 11, RR = 27)

Note that we can find in abstracts many examples of long sequences of words including nouns, ed/ing form verbs, adjectives, adverbs, etc. which form simple noun phrases.

Step 6 Verb Group (main verbs and auxiliaries) (SG = 3, RR = 27)

Step 7 Determination of the Scopes of Coordinated Verb Groups and Clauses (SG = 56, RR = 145)

It is often quite difficult to determine whether the coordination is a clausal one or phrasal one. So, this SG calls recursively the SG for phrasal coordination to check the possibility The two SGs call each other in coordinated ways.

Step 8 Determination of the Scopes of Coordinated Phrases (SG = 29, RR = 199)

Step 9 Appositional and Inserted Phrases/Clauses (SG = 46, RR = 1159)

Step 10 Determination of Global Structures of Sentences and Revision of Tentative Decisions in Preceding Steps (SG = 43, RR = 153)

All the processing up to this step analyze sentences from the bottom to the top. On the other hand, this stage tries to hypothesize the global structures of input sentences in a top-down manner. That is, this step constructs tree structures which show the relationships of clausal constituents in input sentences, while the internal structures of individual clausal constituents have not been determined yet.

This Step first identifies several cue words (or expressions) such as WH words, subordinate conjunctions, 'that', -ing/-ed verbs, to-infinitives, etc. each of which indicates existence of clausal constituents of certain types.

Wrong decision makings in the preceding steps are to be detected and revised by rules referring to these global structures.

Step 11 Disambiguation of Parts of Speech based on Local Environment-2 (SG = 5, RR = 13)

Step 12 Case Pattern Matching (SG = 96, RR = 240)

The scopes of clausal constituents given in the global structures are tentative. The exact scopes are to be determined at the same time when the internal structures of individual clauses are determined.

Step 13 Treatment of Circumstantial Elements (SG = 44, RR = 517)

Semantic interpretation of circumstantial elements is crucial in getting natural translation results.

The interpretations are performed by invoking lexical rules defined for individual prepositions and group prepositions.

Step 14 Conversion from phrase Structures to Dependency Structures (SG = 18, RR = 78)

Step 15 Semantic Interpretation including deep tense, aspect interpretation (SG = 14, RR = 38)

Step 16 Checking Rules of final analysis results (SG = 5, RR = 34)

3.2 Transfer Grammar:

Special care is taken to adjust structural differences of the two languages, English and Japanese, which belong to quite different language families. In order to support global structural changes during the transfer, two separate phases, Pre-Transfer Loop and Post-Transfer Loop, are provided, which are executed before and after the main transfer phase respectively. The detailed construction of the transfer phase and some discussions are given in (5).

- Pre-Transfer Loop: Japanese oriented structures are transformed into more neutral ones (SG = 9, RR = 23)
- Processing of structures governed by predicates (SG = 33, RR = 577)
- Processing of structures governed by nominal concepts (SG = 10, RR = 192)
- Processing of adverbial phrases (SG = 2, RR = 121)
- Processing of Prenominal Modifiers (SG = 21, RR = 180)
- Post-Transfer Loop : The neutral structures are Transformed into English-oriented structures (SG = 11, RR = 46)

3.3 Generation Grammar of Japanese:

- Conversion of parts of speech (SG = 22, RR = 47)
- Selection of Appropriate Case Particles (SG = 16, RR = 103)
- Realization of Voice Information (SG = 28, RR = 34)
- Determination of Phrase Ordering (SG = 33, RR = 44)
- Insertion of Punctuation Marks, etc. (SG = 1, RR = 6)
- Building of Phrase Structure Trees (SG = 79, RR = 160)
- Calculation of inflectional forms for each word and adjustment of structures for morphological generation (SG = 15, RR = 189)

3.4 Evaluation Results:

Detailed evaluation criteria of translation results and results of evaluation are given in (1). About 80 are evaluated as “understandable” by native speakers, which fulfils the initial objectives of the project.

4 Brief Outline of the Mu-II Project

The results of the MU project show economical and technical feasibility of MT in the field of 'translation of abstracts'. The translation results are not so natural as you can expect of those produced by professional (highly qualified) human translators, but they are good enough to get rough ideas about the contents of the papers.

The MU systems as they are, however, have several defects as systems for practical services, because they were initially designed as research and development prototypes. In order to remove certain defects, a new project, the MU-II project, was started. The main objectives of the MU-II are as follows;

4.1 Improvement of Processing Speed and Reduction of Memory Requirement

The new version of GRADE is to be implemented by C in order to improve the processing speed and memory requirements. Some facilities the current version of GRADE provides, which are too flexible and too powerful and which have been rarely used by grammar writers, will be removed. This will also improve the efficiency of processing.

4.2 Augmentation of Dictionaries

The current dictionaries contain about 80,000 lexical items, but they cover only the ordinary words of Japanese, and the vocabularies of certain restricted areas of electrical engineering fields. The dictionaries of the new project are planned to gather 300,000 lexical items including additional ordinary words, terminological expressions in several technological fields.

4.3 Restricted Language

The current version of grammars in the MU systems is designed to translate input sentences whatever strange the inputs are. In fact, we can find many sentences which are difficult even for ordinary native speakers (not the specialists of the fields the input abstracts deal with) to understand. Not a few SGs and RRs are prepared simply for treating very idiosyncratic constructions which even native speakers feel difficulties in understanding. Existence of such SGs and RRs make the systems clumsy and difficult to maintain.

In the new systems, such (almost ungrammatical) sentences are supposed not to be in the text. We will give some trainings to the people who prepare abstracts. However, we keep the principle that MT systems such as systems for translation of abstracts should not rely heavily on pre-editing.

4.4 Other Facilities in the New Systems

Various human interaction facilities are to be integrated around the core systems, including translation oriented text editors, effective utilization of domain specific dictionaries and user oriented dictionaries, maintenance tools for dictionary data bases, etc.

5 Conclusion

The evaluation of the MU project is currently quite positive for further research and development of MT systems. The experiments made with the MU systems show that the basic design principles, especially lexicon driven processings, neutral dictionaries, and heuristically guided processings are extremely effective in the actual development of large scale MT systems. The MU-II project will concentrate on making the current systems practically usable by refining the present softwares and grammars.

However, we also noticed through experiences that there are still many problems, theoretically or practically, which are obviously beyond our current framework. Among others, appropriate target lexical selections, processing of elliptical expressions frequent in Japanese, etc. require a lot further investigations on possible frameworks of semantic and contextual processings in MT. The discussions about integrations of such components is given in (6).

ACKNOWLEDGMENT

I would like to thank my colleagues of the MU project, especially Prof. M. Nagao who is the director of the project and Prof. J. Nakamura and Miss M. Kume who are responsible for softwares and grammars respectively. I also wish to thank Mr. M. Sato and Mr. Y. Sakamoto actually engaged in dictionary development and some other important software development.

REFERENCES

1. Nagao, M., Tsujii, J. and Nakamura, J.: The Japanese Government Project of Machine Translation, Journal of Computational Linguistics, Vol-11, No.2-3, 1985.

2. Nagao, M., et al.: Dealing with Incompleteness of Linguistic Knowledge in Language Translation, in Proc. of Coling84, 1984.

3. Tsujii, J., et al.: Analysis Grammar of Japanese in the MU Machine Translation Systems, in Proc. of Coling 84, 1984.

4. Nakamura, J., et al.: Grammar Writing System (GRADE) of MU-Machine Translation Project and its Characteristics, in Proc. of Coling 84, 1984.

5. Nagao, M., et al.: Transfer Phase of a Machine Translation system, in Proc. of Coling 86, 1986.

6. Tsujii, J.: Future Direction of Machine Translation, in Proc. of Coling 86, 1986.