[MT Summit IV, July 20-22, 1993, Kobe, Japan]

Evaluation Method of Machine Translation: From the Viewpoint of Natural Language Processing

Strategy for Machine Translation Evaluation

Hirosato Nomura
Dept. of Artificial Intelligence
Kyushu Institute of Technology
Iizuka, 820, Japan

1. Introduction

This is briefly introducing experiences on works of machine translation evaluation and additional personal considerations on the matter. It is arranged by the issues provided by the panel organizer.

2. Methodology

Discussion of machine translation evaluation has to be approached from several stand points. The first point is on technology concerning multi-lingual computer processing. It relates to all issues on designing and developing actual machine translation systems based on the current technology, thus involving issues on lexicon, grammar, term bank, parsing and generation, system integration, human interface, etc. It is to check whether available data and technology have been well encoded into the MT system and then is well functioning on it so that it can provide the most powerful but economic MT system to each user. The second point concerns a guideline for user selection of a MT system among candidates when purchasing, by which the user can expect the most cost saving translation business. The third point is to find future technical problems on machine translation, which will help researchers and developers to recognize what tasks to be elaborated. From these points of view, we at the JEIDA MT Committee have been discussing and studying Machine Translation Evaluation in these several years, and proposed the first version of the evaluation criteria in 1991 and its revised version in 1993. Several reports and technical papers have been published as listed at the end of this paper. The first version of the criteria is divided into two parts: User evaluation, Developer Technological evaluation. The revised version consists of three parts: User evaluation, Developer evaluation and Quality evaluation. For the quality evaluation we have been designing test corpus for checking translation accuracy. We will also list technical items and their scoring criteria to be specified from the test corpus. We have been simulating our criteria in the field and we have already got a lot of useful and interesting reactions from it. Those data accumulated through the experiences will be considered to incorporate into the further revision of our evaluation criteria.

3. Strategies

The state of the art will determine what kind of evaluation has to be elaborated. If the stage were in producing grammatically correct translation, then the purpose of the evaluation will concern naturalness or goodness of the produced translations as a whole text. However, if the stage were in discussing how to design MT system which can translate 80% of sentences, then the purpose of the evaluation should concentrate on linguistic encoding. Anyway, linguistic phenomena is enough complicated to encode detail linguistic information thus the most significant but fundamental problem is to provide enough data so that the system can produce grammatically acceptable translation, otherwise any other ambitious challenges will bring nothing at all. Since MT involves variety of difficult problems even in the area of natural language processing, it is very natural that the evaluation concerns issues which cover all of the problems. Thus, simpleness and plainness of the criteria will be the secondary problem to be concerned when discussing the methodology of MT evaluation while we have been trying to design the simpler and easier criteria as possible. For the purpose, we developed a check list as a collection of items to be evaluated, on which checking are conducted one by one. Each item states each problem in isolated and simple manner thus can be easily checked. The scoring of each check is managed objectively thus very easy. The visual radar chart format is provided, which is very friendly to the evaluators when making the final judgment.

4. Problems

As mentioned earlier, we have concerned three aspects of machine translation evaluation: user evaluation, developer evaluation, and quality evaluation, each of which relates to marketing of MT system, future development, and current status. These three aspects can be compared with the three categories in EC evaluation methodology. It is obvious that these three points have to be discussed separately (but not independently) since users should have chances to select the most suitable MT system satisfying their purposes and situations, developers should recognize the current state of the art of MT technology, and researchers/engineers should study what to and how to develop more advanced MT systems which might be able to meet wider variety of current and future demands. To make clear each problem and their relationships, we have been developing a hierarchical structure of items which will turn to a thesaurus and knowledge base for terms on machine translation.

5. Standard and Cooperation

In order to achieve the cost saving, efficient and reliable evaluation, it is better to provide some commonly usable or standardized test corpus and evaluation algorithm, which is quite possible if we do not expect the perfect one. Such a test corpus must involve translation equivalence for each source sentence as well as a lot of linguistic information which will be applied for evaluation. Such tasks for providing multi-lingual text corpus and multi-lingual linguistic information are not easy thus much elaboration as well as international cooperation will be invoked. However, it must be understood that the current MT technology is not necessarily well developed thus is still on the way to the goal. The fact reveals that the development of the test corpus and evaluation algorithm must be promoted through cut and error manner. Those two axes, one for technological consideration and the other for standardization and international cooperation, have to be collaborated jointly.

References (Listing only related ones to this paper):

1) JEIDA, A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, U.S.A., JEIDA MT Committee, 1989

2) Nomura and Isahara, JEIDA's criteria on machine translation evaluation, Proc. of Int. Sym. on NLU& AI, July 1992

3) Nomura and Isahara, JEIDA methodology and criteria on machine translation evaluation, a book delivered at the Sandiego Workshop on Machine Translation Evaluation Workshop, Nov. 1992

4) JEIDA, The survey of the current status of research and future trends in machine translation and natural language processing, JEIDA MT Committee, Dec. 1992

5) JEIDA, The revised version of JEIDA criteria on machine translation evaluation, 1992 JEIDA report of the MT committee, March 1993 (in Japanese)

6) Takayama, Itoh, Yagisawa, Mogi and Nomura, JEIDA's proposed method for evaluating machine translation (End user system selection), Proc. of SIGNLP 93-NL-96, Sept. 1993 (in Japanese)

7) Nakaiwa, Morimoto, Matsudaira, Narita and Nomura, JEIDA's proposed method for evaluating machine translation (Developer's guidelines), Proc. of SIGNLP 93-NL-96, Sept. 1993 (in Japanese)

8) Isahara, Shin-nou, Yamabana, Moriguchi and Nomura, JEIDA's proposed method for evaluating machine translation (Translation quality), Proc. of SIGNLP 93-NL-96, Sept. 1993 (in Japanese)