

## **EVALUATION SURVEYS: The JEIDA Methodology and Survey**

*Hirosato Nomura*

Kyushu Institute of Technology

*Hitoshi Isahara*

Electrotechnical Laboratory

We have developed two sets of criteria for machine translation evaluation: one by which users evaluate the possibility of introducing a machine translation system, and the other by which researchers and developers undertake technical evaluations of their systems for future research and development.

There are two types of criteria for users: economic criteria and technical criteria. The set of economic criteria for users was designed to determine which type of machine translation system should be introduced in order to obtain maximum economic efficiency. The technical set of criteria for users was designed to help those who have decided to introduce a machine translation system to determine which system will best satisfy their needs.

With regard to the criteria for developers, the technical aspects involved in machine translation are listed hierarchically, while focusing on the relationship between them. Criteria are provided for evaluating the level of achievement of each technology as well as estimating the future tasks to be developed in the respective areas. The machine translation evaluation criteria proposed here are designed so that the evaluation of each item can be made objectively (objective ratings); so that the overall evaluation of each item can be done numerically (numerical ratings); and so that the final clarification can be as visual as possible (visual judgment).

### **1. Economic criteria for users**

This set of criteria proposes the ideal type of machine translation system, based on economic factors, for users who intend to introduce such a system. Each set proposes two types of machine translation systems: one is the type of machine translation system most suitable for the user's present situation, and the other is the type of system which will best satisfy the user's needs. More often than not, the two types of systems proposed are identical. If not, users can recognize their current problems or the improvements to be made on their operation if a machine translation system is introduced. This means that this set of criteria aims to support the decision-making of the users who are planning to introduce a machine translation system as soon as possible or in the near future.

The users answer two questionnaires, i.e., one to determine the user's present situation, and the other to determine the user's needs, to make radar charts for them. These radar charts represent 14 parameters characterizing the machine translation system. The evaluation value for each parameter varies depending on the answers to the questions. In the meantime, machine translation systems are classified into seven types, and the user considered most suitable for each system are assumed. Radar charts for these system types are then drawn on the basis of the answers the user is expected to give on the questionnaire. The users choose the most suitable type of machine translation system by comparing the radar charts for the seven system types with the two radar charts, one for the user's present situation, and the other for the user's needs.

## 2. Technical Criteria for Users

This criterion aims to assist users who plan to introduce a machine translation system. The main purpose of this criterion is to evaluate “the degree of the user’s satisfaction with a machine translation system” in comparison with the user’s needs, rather than to evaluate the machine translation system itself. When introducing a machine translation system, it is necessary to evaluate various factors, such as the quality of translation, introduction costs, maintenance costs, pre-editing and postediting functions, and the quality of the dictionary. These factors vary in importance depending on how the system is used. Therefore, when technical features are evaluated, it is necessary to change the factors to be focused on while at the same time not using sight of the user’s needs.

Under this heading there are two questionnaires, i.e., (A) the questionnaire to clarify the user’s requirement for machine translation, and (B) the questionnaire to clarify the overall functions of a machine translation system. After completing questionnaires (A) and (B), the user can determine whether or not the system meets the users’ requirements. Here, this set of criteria provides the score for each aspect of evaluation (31 parameters in total) while referring to questionnaires A and B.

These scores are used to calculate the score for 10 axes on a radar chart which is the output of evaluation by this criterion. This is a visual graph which represents both the system performance and the degree of the user’s satisfaction with the system. This evaluation is made with 10-point scoring for 10 axes: What Is to Be Done?, Speed of Translation, Quality of Translation without Adjustment, Quality of Translation after Adjustment, Ability to Edit, System Configuration, Style of Operation, Personnel, Introduction of System, and System. The user then analyzes the radar charts for several systems and chooses the most suitable system.

The advantage of this set of criteria is the ability to evaluate the degree of the user’s satisfaction with the machine translation system by comparing the user’s needs with the system specification.

## 3. Criteria for Developers

This set of criteria aims to assist researchers and developers of a machine translation system when evaluating the system they have developed. It is intended for making an in-house evaluation of the technical level the system has achieved and whether the system suits the purpose of development. When a machine translation system is evaluated from the developer’s point of view, there seem to be many evaluation items. The following aspects are covered: (1) Features of the System (2) Dictionary (3) Analysis (4) Intermediate Representation and Processing (5) Generation (6) Processing of Special Forms (7) Customization and Learning Function (8) Environment and Operation.

This set of criteria permit both the overall evaluation of the system and the evaluation of technical components which have been incorporated into the system. As a result, some advantages or disadvantages of the system, either general or specific, can be recognized by the developer. However, this set of criteria does not evaluate the quality of machine translation system explicitly.

We are also developing a new set of criteria to more precisely evaluate the quality of the translation itself. The Machine Translation Market and Technology Study Committee (Prof. Hirosato Nomura, Chairman) of JEIDA (Japan Electronic Industry Development Association) has been working to develop criteria for machine translation evaluation for several years and has now completed the first version. A 300-page report was published on it in March 1992.

The paper presents the JEIDA discussion on the methodology and then lists the criteria.