Technical Evaluation of MT Systems from the Developer's Point of View: Exploiting Test-Sets for Quality Evaluation

Isahara, H.*(Electrotechnical Laboratory, MITI),
Uchino, H. (Nippon Telegraph and Telephone), Ogino, S. (Nihon IBM),
Okunishi, T. (Sharp Corp.), Kinoshita, S. (Toshiba),
Shibata, S. (CANON INC.), Sugio, T. (Oki Electric Industry Co., Ltd.),
Takayama, Y. (Mitsubishi Electric Corp.), Doi, S. (NEC Corp.),
Nagano, T. (Matsushita Electric Industrial Co., Ltd.), Narita, M. (Ricoh Co., Ltd.),
and Nomura, H. (Kyushu Institute of Technology)

Abstract

This paper describes a method of evaluating quality for developers of machine translation systems to easily check imperfections in their own systems. This evaluation method is a systematic, objective method along with test example sets in which we clarified the evaluation procedure by adding yes/no questions and explanations to the example sentences for evaluation.

1 Introduction

Since 1992, we have been developing a method of evaluating quality for the developers of machine translation (MT) systems to easily check imperfections in their own systems [1, 2, and 3]. In this paper, we would like to describe this systematic, objective method along with the test example sets in which we have clarified the evaluation procedure by adding questions and explanations to the examples for the evaluation[1].

We will first describe how our evaluation method surpasses previous methods, with reference to the following 2 types of objectivity.

(1) Objectivity in the evaluation process

(2) Objectivity in judgment of the evaluation results

---

*isahara@etl.go.jp

[1] The work described in this paper is being developed by the Special Interest Group on Machine Translation (Chief: Hitoshi ISAHARA, Electrotechnical Laboratory) in the Natural Language Processing System Research Committee (Chairman: Prof. Hozumi TANAKA, Tokyo Institute of Technology) which is a subcommittee of the Natural Language Processing Technology Committee (Chairman: Prof. Makoto NAGAO, Kyoto University) of JEIDA (Japan Electronic Industry Development Association). JEIDA has formulated three criteria for evaluating MT systems: 1) technical and 2) economical evaluations for system users, and 3) technical evaluation for the system developers. For more information on these criteria, please refer to references 1 and 4.

In an evaluation method such as the one proposed in the ALPAC report, "fidelity" and "intelligibility" are employed as evaluation measures, though they are dependent on human, subjective judgment. Consequently, the results may differ according to who has made the evaluations, that is, they do not satisfy the objectivity criterion (1). Theoretically, the evaluation method in the ALPAC report satisfies criterion (2) since the evaluation results are given as numbers. The system developers, however, fail to recognize which items cannot be handled in their own system. This is because the test example in question covers various kinds of grammatical items. So, their interpretation of the evaluation result for further improvement of their system must still be subjective. Therefore, for all practical purposes, this evaluation method does not satisfy criterion (2).

On the other hand, we have been preparing test-sets that can satisfy both objectivity criteria (1) and (2). There, we have clarified how to evaluate individual examples posing yes/no questions which enable the system developers to make an evaluation just by answering them. With our method, everyone can evaluate MT systems equally, for his/her answer requires only a simple yes or no. Even for imperfect translation results, judgment will not vary widely among evaluators. In addition, we have assigned to each example an explanation which gives the relationship of the translation mechanism to the linguistic phenomenon, thus enabling the system developer to know why the linguistic phenomenon in question was not analyzed correctly. Consequently, with our test-set method, the evaluation results can be utilized for improving MT systems.

There is another proposed method where example evaluation sentences are collected. Each example sentence relates to a linguistic phenomenon subject to evaluation [5, 6, and 7]. With these test-sets, if a system is evaluated as incapable of properly translating an example, the system developer can immediately recognize that his/her system cannot handle the linguistic phenomenon in question. Therefore, we can conclude that this method satisfies the objectivity criterion (2). At present, however, this method has the following two problems:

(1) The procedure for evaluating the translation output has not been clarified.

(2) Learning deficiencies of the MT system via the evaluation results is dependent on the linguistic intuition of the evaluator.

As long as it is based on the example sentences simply collected as the test-sets, this method can be used for ad hoc evaluation only, and cannot be established as an evaluation method. Moreover, to enable evaluation results to be used for improving MT systems, the listing of various linguistic phenomena is not enough; it is also necessary to clarify the positioning of each linguistic phenomenon within the grammar.

In our test-sets, we have systematically sampled the grammatical items that ought to be taken up, and listed some examples for each item. The test-sets clearly describe what linguistic phenomenon should be evaluated in each example so that the developers can easily understand the problems they need to solve in their systems. The system developer can identify causes of translation failures.

In Chapter 2, we will describe our method of quality evaluation, i.e., what information should be provided to system developers as a result of quality evaluation of MT systems. Chapter 3 describes how the test examples were collected and should be evaluated. Chapters 4 and 5 give some examples to show the test-sets for English-to-Japanese MT systems and Japanese-to-English MT systems, respectively.

2 Standpoint for the Evaluation Method

The method we propose here is a quality evaluation method which is totally independent of the MT system design. Therefore, the system developer can use this method regardless of his/her

system type, i.e., whether the relevant MT system is rule-based or example-based. Conversely, in this method, if it becomes clear that a specific linguistic phenomenon cannot be processed on the relevant MT system, no solution common to the various system types is indicated, so the solution is entrusted to the developer according to the specific system type.

In our test-sets, we give no information on how often the linguistic phenomenon in each test-set appears in general usage. This is because the frequency of appearance of the relevant linguistic phenomenon might differ according to the type of document to be translated. If specific linguistic phenomena regularly appear in the documents handled on a specific MT system, the evaluator needs only to select the test-set which corresponds to the linguistic phenomena in question. Wrong evaluations could be made if scoring was based merely on the frequency of individual linguistic phenomenon.

To sum up, this evaluation method is designed in such a way that the system developers, irrespective of their system type, can precisely understand linguistic phenomena which cannot be handled by their systems and thus should be taken into account when improving the system performance.

3 Characteristics of the Test-Sets for Quality Evaluation

The test-sets employed in our evaluation method consist of example sentences for evaluation, their model translations (human translations), and the questions by which MT outputs should be evaluated. With the test-sets, the MT system developers can make objective judgments on the translation quality just by preparing the system output and answering the question assigned to each example sentence. This chapter describes how the example sentences were collected for the test-sets, and how the actual evaluation is made using the test-sets.

3.1 Collection of Example Sentences for Evaluation

The example sentences in the test-sets were collected by researchers and engineers who have actually dealt with the development of MT systems and/or natural language processing systems. During the collection of the examples, we emphasized the following two points:

(1) Coverage of all the basic linguistic phenomena

(2) Selection of examples with linguistic phenomena that are difficult to handle with MT systems, especially those with ambiguity problems

In other words, (1) refers to a systematic specification of the grammatical phenomena to be evaluated (top-down approach) and collecting examples according to these phenomena. On the other hand, (2) refers to a collection of examples that are difficult to translate on MT systems (bottom-up approach). In particular, we concentrated on those linguistic phenomena whose processing difficulties may be solved in the near future. Then, we systematized the examples for evaluation of MT systems. Furthermore, we repeated the translation evaluation tests on those examples using some commercial systems, and improved the test-sets focusing on the following points. All of them are important factors for maintaining objectivity during the evaluation process.

- No ambiguity in the questions
- No unnecessary complexity in any example
- No ambiguity in the translation of any example

```
◎Sentaku-Kobun-Hinsi-Noun/aux
(= Select - Sentence Structure - Part of Speech - Noun/auxiliary verb)
【番号】  189   (= 【ID No.】)
【例文】  The trash can was thrown away.   (= 【Example】)
【訳文】  ごみカンは捨てられた。   (= 【Translation】)
【質問】  "can" が「カン / 缶」のように名詞として訳されていますか？
          (= 【Q.】 Is "can" translated as a noun?)
【訳出例】○ (くず缶 / ごみ容器 / くず入れ) は (廃棄された /[投げ] 捨てられた)。
          × ごみは捨てられ得る。
          (= 【Translation Samples】 literally meaning:
                yes: The (garbage can/trash bin/litter bin) was (discarded/[thrown]
                     dumped).
                no: The trash can be discarded. )
【類似文】The last will was opened.「最後の遺言書は開けられた。」
          (= 【Similar Examples】 and the Japanese translation)
【関連文番号】190, 191   (= 【Related ID No.】)
【解説】  "can was" の並びから、"can" が助動詞でないことがわかる。
          (= 【Explanation】 The word order of "can was" shows that "can" is not an
                             auxiliary verb.)

【番号】  190   (= 【ID No.】)
【例文】  The trash can be thrown away.   (= 【Example】)
【訳文】  ごみは捨てられ得る。   (= 【Translation】)
【質問】  "can" が「～できる / 得る」のように助動詞として訳されていますか？
          (= 【Q.】 Is "can" translated as an auxiliary verb meaning "has ability to/has a
                    possibility to"?)
【訳出例】○ (くず / ごみ) は (廃棄できる /[投げ] 捨てられることができる)。
          × ごみカンは捨てられた。
          (= 【Translation Samples】 literally meaning:
                yes: The (garbage/trash) (can be discarded/[thrown] dumped).
                no: The trash can was discarded. )
【類似文】(= 【Similar Examples】)
【関連文番号】189, 191   (= 【Related ID No.】)
【解説】  189 とは逆に、ここでは "can" は名詞ではなく助動詞。
          (= 【Explanation】 In contrast to No.189, here, "can" is not a noun but an
                             auxiliary verb.)
```

Fig. 1 Sample Test-Sets for English-to-Japanese MT systems

3.2 Evaluation Method

Evaluation of the translation results is conducted as follows.

(1) Translating the example sentences in the test-sets with MT systems

(2) Checking the translation results of (1), and answering each example's individual question

We specified the judging points in the questions (e.g. which part of the example plays the grammatical role in question, and how that part should be translated), and we posed the questions in a yes/no style, thus avoiding varying judgments among the evaluators. Moreover, sample answers were also assigned to each test-set which were based on the translation results of five types of existing

commercial MT systems (at present, in "the Test-Sets for English-to-Japanese MT Systems" only). [1] By referring to them, judgment can be easily made on each question.

4 Test-Sets for English-to-Japanese MT Systems

As the initial step in constructing the test example sets for English-to-Japanese MT systems, we selected mainly simple English sentences as test items. We studied and evaluated 309 examples of basic sentences, and compiled them in our "1993 Test Sets." Our current work is to extend the test examples to complex sentences. The test-sets will be entirely completed by the end of March, 1995. We have also been evaluating the test-sets with 5 different English-to-Japanese MT systems in order to examine their practicability, rewriting the questions in the test-sets if necessary.

Each test-set consists of: an ID number, an example, a model translation, a yes/no question, translation sample(s) by MT systems, a sentence or sentences with similar syntax, ID number(s) of the related example(s), and explanation (See Fig. 1). In this chapter, the Quality Evaluation Process, Object's Linguistic Phenomena, and the Simulation on MT systems are described.

4.1 Evaluation Process

Evaluation of the quality of English-to-Japanese MT systems is conducted as follows.

- To translate [Example] in each test-set with English-to-Japanese MT systems.
- To answer "yes" or "no" (O or X) to a question on each example by referring to the translation result.
- To check the distribution of "yes'" and "no's" in the test-sets and evaluate the system performance.

With the yes/no distribution, the system developer can easily pinpoint the items which his/her system did not translate properly. In the test-sets, however, differences in significance and frequency among the examples are not taken into consideration. Therefore, it is meaningless to simply count the number of "yes" answers to compare the performance of various MT systems.

4.2 Linguistic Phenomena as Test Object

The test-sets consist of 309 basic, mainly simple English sentences as follows.

```
Structural Analysis Part
     Sentence Pattern:          55 examples
     Temporal Information:      63 examples
     Auxiliary Verbs:           37 examples
    Sentence Type:               29 examples
      --Subtotal:               184 examples
 Structural Selection Part
   Structural Disambiguation: 56 examples
   Semantic Disambiguation (by co-occurring word(s)):
                               69 examples
     —Subtotal:               125 examples
     —TOTAL:                  309 examples
```

As shown above, the quality evaluation items were collected from the following perspectives: "Structural Analysis" and "Structural Selection."

In the "Analysis" part, MT systems are checked as to whether they can correctly analyze the sentence structure of the test example. This is a top-down approach in which the comprehensiveness of MT systems is checked. In a word, this part is intended to judge whether the MT system in question meets the requirements for an MT with good performance. Grammatical phenomena essential for English were classified into 4 groups, referring to some grammar books (see [8,9 and 10]): (1) Sentence Pattern, (2) Temporal Information, (3) Auxiliary Verbs and (4) Sentence Type. Sentence Patterns were selected based on Hornby's classification. In doing so, some patterns were intentionally omitted because they were judged to be unnecessary for quality evaluation of MT systems. In addition, some usages of auxiliary verbs were omitted because they were considered to rarely appear in the documents for MT systems.

In the "Selection" part, on the other hand, MT systems are checked as to whether they can identify the correct structure syntactically and/or semantically when example sentences provide ambiguity problems. This is a bottom-up approach in which the disambiguating ability of MT systems should be checked. Thus example sentences were classified into two groups: (1) Structural Disambiguation and (2) Semantic Disambiguation.

4.3 Test-Set Simulation on MT Systems

In order to examine the practicability of the test-sets, we conducted a translation simulation on the five MT systems. The correct answer rates of the five systems differed greatly: from 53 to 80 percent. Though these rates alone do not have any significance, they do indicate that the five systems are quite different in performance both in the "Analysis" part and in the "Selection" part. That is to say, our test-sets have successfully revealed that the range of linguistic phenomena which each MT system can handle is quite different. Therefore, the method that we have proposed here allows an efficient quality evaluation of MT systems.

5 Test-Sets for Japanese-to-English MT Systems

In order to evaluate the ability of Japanese-to-English MT systems, two kinds of proposals have been made so far. The first one focused on the difference in the way of perception between English-speaking people and Japanese people and thus classified Japanese expressions so that they can be used as test examples [5 and 6]. On the other hand, the second one focused on the structure of Japanese expressions and proposed example sentences for evaluation which typically represent the structural characteristics of Japanese expressions [7].

Our test-sets for Japanese-to-English MT systems began to be constructed in 1993. Like those for English-to-Japanese MT systems, they are intended to clarify what is insufficient in their systems by answering the questions. However, we have constructed the test-sets for Japanese-to-English MT systems from a slightly different perspective than we have done for English-to-Japanese MT systems. Fig. 2 shows a sample test-set for Japanese-to-English MT systems.

In our approach, we have not only employed the test-sets which enables an objective evaluation of MT systems but also established an evaluation method which enables the developers of Japanese processing systems to identify the correspondence between the linguistic phenomena and the processing modules. That is to say, in addition to example sentences and their evaluation procedure, questions have been assigned to each test-set so that the evaluator can check how his/her system handles the linguistic phenomenon in question. In this way, the system developer can evaluate

(1-4)Complex Predicates

【Explanation】 In order to translate a complex predicate, the grammatical relation between the components, i.e., the complement-verb combination or adjunct-verb combination, should be accurately identified.

[Identification of the Complement-Verb Combination]
Question:How can this relation be identified in your system?
☐ By using the semantic constraint on the word which can function as a complement of the verb
☐ By judging the verbal component to determine the default grammatical relation
☐ Other methods (　　)
Example Sentence: 住民が自然保護する
Complex Predicate: 自然保護する
Model Translation: The residents conserve nature.
Identification: The word 「自然」 should be identified as the object of the verb 「保護する」.

[Identification of the Adjunct-Verb Combination]
Question:How can this relation be identified in your system?
☐ By identifying the word which can function as an adjunct of the verb based on the word type
☐ By examining the possibility of co-occurrence between the adjunct candidate and the verb
☐ Other methods (　　)
Example Sentence: 渋滞が自然解消する
Complex Predicate: 自然解消する
Model Translation: The traffic jam dissolved naturally.
Identification: The word「自然」should be identified as adverbial adjunct of the verb 「解消する」.

【Comments】 The complex predicates「自然保護する」and「自然解消する」include the same word 「自然」that has a different grammatical function in each predicate. In the former, this word functions as the complement of the verb「保護する」, while functioning as an adverbial adjunct of the verb「解消する」in the latter.

Fig. 2　Sample of the Test-Sets for Japanese-to-English MT Systems

the processing ability of his/her system as a whole and also recognize the performance of each processing module of his/her system.

In our test-sets, linguistic phenomena in Japanese were classified into 40 categories. To each category, a question has been given so as to check how the linguistic phenomenon in question is handled. If necessary, additional questions have been assigned to clarify the knowledge in use and how to deal with the output of the process. Each linguistic phenomenon is exemplified in test sentences and provided with a model translation in English and an explanation about the key factors in translation. So far, 350 technical sentences have been selected as test sentences. These test sentences are currently under examination via translation experiments with commercial MT software. Explanations described in the test-sets are also to be modified.

Moreover, a check list is available in our test-sets. This list can be used by the system developers to check the correspondence between the linguistic phenomenon in question and the processing

module to be engaged in handling it. This makes it possible to judge which processing module is responsible for the inadequacy of the system output. It is also possible for the evaluator to modify this check list.

Our test-sets for Japanese-to-English MT systems will be completed by March, 1995, along with those for English-to-Japanese MT systems.

6 Conclusion

In this paper, we have proposed systematic and objective methods for evaluating the translation quality of the MT system from the developer's point of view.

Our method employs test-sets in which example sentences, their model translations, questions for evaluating the system output, similar examples (if any), and grammatical explanations have been systematically aligned. The example sentences have been collected focusing on wide coverage of (1) basic linguistic phenomena and (2) linguistic phenomena problematic to MT systems.

The questions in the test-sets are designed to clarify the evaluation viewpoints. Given the system outputs corresponding to the example sentence in question, the system developer needs only to answer the question assigned to the example sentence. This judgment does not vary among the evaluators, thus enabling an objective evaluation. Furthermore, with our test-sets, the system developer can precisely recognize which linguistic phenomena cannot be handled by his/her own system.

Our future plans are (1) to solve existing problems revealed by evaluation experiments with some commercial MT systems and (2) to increase the number of example sentences so as to cover more linguistic phenomena. When our two kinds of test-sets (English-to-Japanese and Japanese-to-English test-sets) are completed next March, we will make them available to the public. Finally, we hope our evaluation method can play a useful role in the development of MT systems.

References

[1] "Survey Report on Machine Translation Systems" (in Japanese), Japan Electronic Industry Development Association (JEIDA), 1993.

[2] "Survey Report on the Natural Language Processing Technology" (in Japanese), Japan Electronic Industry Development Association (JEIDA), 1994.

[3] H. Isahara, et al. : "JEIDA's Proposed Method for Evaluating Machine Translation (Translation Quality) — A Proposed Standard Method and Corpus —" (in Japanese), IPSJ SIG Report, NL96-11 ,1993.

[4] H. Nomura and H. Isahara : "JEIDA's Criteria on Machine Translation Evaluation", International Symposium on Natural Language Understanding and AI, 1992.

[5] S. Ikehara and S. Shirai : "Function Test System for Japanese to English Machine Translation" (in Japanese), IEICE SIG Report, NLC90-43, 1990.

[6] S. Ikehara et al. : "Criteria for Evaluating the Linguistic Quality of Japanese to English Machine Translations" (in Japanese), J. of Japanese Society for Artificial Intelligence, Vol. 9, No. 4, 1994.

[7] H. Narita : "An Criteria of Processing Ability for Sentence Structure" (in Japanese), IPSJ SIG Report, NL69-1, 1988.

[8] A. S. Hornby : "Guide to Patterns and Usage in English, Second edition", Oxford Univ. Press, 1975.

[9] Y. Ogawa, et al. : "The Wonder Book of English Grammar" (in Japanese), Obun-sha , Tokyo, 1991.

[10] Y. Egawa : "Explanation on the English Grammar" (in Japanese), Kaneko Shobo, 1964.