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

Shoichi Yokoyama
Department of Electrical and Information Engineering
Faculty of Engineering, Yamagata University
4-3-16, Jonan, Yonezawa, Yamagata 992, Japan
Tel. +81-238-22-5181 ext.~445, FAX +81-238-22-8009
e-mail: yokoyama@emtsun.yz.yamagata-u.ac.jp

1. Summarization of My Experience

I am a vice-chairperson of the System Evaluation Working Group (chairperson is Prof. Hozumi TANAKA, Tokyo Institute of Technology) in AAMT (Asia-Pacific Association for Machine Translation). The working group was started in 1991. Now 16 members attend a monthly meeting, and discuss the issues of evaluation of machine translation. In the following paragraphs, I simply introduce the standpoint and purpose of the working group.

Recently, various machine translation system have been used for machine assisted human translation in Japan. There have been some trials for evaluating these machine translation systems. In general, human checks the output sentences and gives them a score for faithfulness and understandability. In some other trials, a checking list is given to users, and the evaluation rate is the standard used for adopting the machine translation system.

We considered the evaluation of machine translation system from the viewpoint of a natural language processing system (Yokoyama92).

In a machine translation system, almost all of the various processing is based on the natural language processing concepts.

Practically, we list up the sentences difficult to machine-translate. About 100 sentences are listed from both the makers' and users' side. We classify these sentences based on the report from JEIDA, but some other terms are added. These sentences are machine-translated without pre-processing. We analyse input and output sentences and re-classify them according to their difficulty. We also consider alternative sentences by humans. These sentences are also machine-translated.

We analyze the set of sentences and their translations. From these translations we divide the problems as occurring in one of three stages used in usual machine translation systems, that is, (Japanese) analysis phase, transfer phase, and synthesis phase. We deal with a Japanese-English machine translation system this time because deriving the problems of Japanese analysis is a good topic for us, being native Japanese speakers.

Our purposes are as follows:

- We classify and analyze sentences difficult to machine-translate, and from this derive the key issues that need to be resolved for the construction of machine translation systems.

- We consider the reason why translation of these sentences leads to error, and derive from this issues for natural language processing.

- If we succeed in deriving a general resolution of these issues, we establish a new idea for machine translation.

- We collect such sentences, and we will make these fundamental test-beds for examination of the capabilities of machine translation.

- We analyze the transfer part by comparing bilingual lexicons, and we improve the transfer lexicons.

- We create the alternative sentences to original ones, and these sentences will be also treated as candidates for testing the capabilities of machine translation.

Our study continues and (we hope) the analysis will be advanced in detail.

I have been studying natural language processing for Japanese, but had no experience about the study of machine translation system because I believe that my system could not overcome my own poor ability of foreign languages. However, everyone could become a critic even if he could not write novels.

2. About the Claim of a Recent Report

My answer is "yes" and "no." I agree with the opinion that "it is (I don't think quite) unreasonable to look for common, or simple, evaluation techniques ..." However, I don't agree with the "analogy with training the cook." Even with cooking, there are many cooking books or receipts, and one can well cook following their advice. So, not so simple, but good methodology of the evaluation for the machine translation will be possible. In the first stage, there is the limitation of fields and/or vocabulary, but continuous improvement makes the evaluation quality well and common.

3. About the EAGLES Group on Evaluation

"Progress evaluation" is an interesting method. It is very well that the progress of a system goes with the progress of an evaluation. However, it is difficult to evaluate the rate of the progress evaluation, and it would be necessary to evaluate the evaluation, that is, meta-evaluation. Can the rate of the progress be measured automatically and objectively, or by human? That is an attractive and interesting issue, but the difficulty exists.

I don't accept the distinction between different types of evaluation because each evaluation include other evaluations, and the clear distinction is difficult.

4. Cost of Providing Test Materials

I think that it is possible to establish a standard set of sentences. So, for reducing the cost, the test materials across evaluations can be shared. However, it is difficult to share them across a number of evaluations. If the evaluation changes with the progress, the materials such as corpora, vocabulary must change, but as mentioned above, it is quite difficult.

5. Free Comment

I will comment on other topics at the panel discussion for want of the space.

Bibliography

1. Yokoyama92: Shoichi Yokoyama: Toward a Systematic Evaluation of Machine Translation: from the Viewpoint of Natural Language Processing, Proceedings of International Symposium on Natural Language Understanding and AI as a Part of International Symposia on Information Sciences (ISKIT'92),} pp.102-106, 1992.