A Method of Evaluation of the Quality of Translated Text

Yamauchi Satoshi
Research and Development Group
RICOH COMPANY. LTD.
16-1 Shinei-cho Tsuzuki-ku Yokohama, 224-0035, JAPAN
yamauchi@int.rdc.ricoh.co.jp

Abstract

In this paper, I present a method for the evaluation of the quality of translated text, namely, a translation ability index, which shows the relative position of the translation ability of a Machine Translation (MT) system on a measurement scale. The measurements are made by an analysis ratio which is able to make absolute measurements and a conversion and magnitude scale (CGMS) which indicates the mutual relation of the machine translated text to the text translated by the professional human translator. The translation ability index in this work has been confirmed by the evaluation of two MT systems. This is effective as a clear explanation of this work.

1 Introduction

Evaluating or measuring techniques are very important for promoting and indicating progress made in technologies in general. In Japan, the MT systems have been used by a certain number of people in different limited domains. However, anyone who has ever used a MT system hopes that it will be improved upon. So, we have to make a breakthrough in MT technology, and build up an evaluation method that is easy and effective for people to use in choosing a MT system.

After ALPAC (Automatic Language Processing Advisory Committee) [1] had first introduced an evaluation method for the quality of translated text in 1966, several evaluation methods were offered from many researchers and research groups [2][3][4][5]. No method, however, was objective but each one had a partly a subjective factor since two kinds of text belonging to two different cultures must be compared. When an individual evaluates a MT system, his or her judgment when using an evaluation criteria is not consistent. So the result of the evaluation may not be unique because the person who evaluates each translated text has to choose a ranking from various steps determined by the evaluation criteria.

I introduce a new evaluation method that makes it easy for users to choose a MT system without their worrying about issues of the quality of the translated text. The heart of the method is to evaluate generated text in the target language in all translation procedures. There are two steps of evaluation in the total translation ability evaluation. The first step of the evaluation is to check the source text analysis ability. The next step is to check for quality of conversion and generation. The product of these results is the translation ability index. In the following description, I show the result which was obtained in the evaluation of two MT systems.

2 The Evaluation Method

For the evaluation, you need three sets of documents as part of the evaluation criteria. (The original texts that were translated were selected at random in a limited area or field, ignoring any criteria such as suitability for translation by computer.) Bilingual documents must be prepared, i.e. the original texts and their translations. The first set must be completely correct and acceptable in each sub-language (the gold standard). For the second set. the human translator produces translated documents. which mimic documents produced by an ideal MT system, that is, they still need to be corrected for cultural applicability. The third set contains the translated documents which are produced by a machine translation system from the original texts.

In order to understand the following explanations easily, the several sets of texts or documents are named as follows. The first set contains sentences translated by Humans, which is assigned the symbol H. The second set contains Indirectly translated sentences or Ideal MT system translated sentences, which is assigned the symbol I. The third set is the set produced by the MT system, that is, the set of sentences translated by the target MT system. This is the set of correct sentences chosen in the analysis phase, and it is assigned the symbol S.

2.1 The Source-Text Analysis Ratio

In this evaluation step, the evaluator analyzes sentences translated by the MT system, by comparing them with the gold standard (the sentences previously translated by humans). The evaluator must meet at least the following conditions: he or she must be a native speaker of the target language and must understand the grammar of the source language. Having a knowledge of the domain area is useful.

After the system being evaluated has translated the original documents, then the evaluator selects the syntactically correct sentences from among the translated sentences (errors from similar words with different semantic meaning are still present after this evaluation step). The analysis rate of the source language is the proportion of

$$M_A = \frac{C_{sent.}}{T_{sent.}} \times 100 \ (\%) \qquad \text{.........} (1)$$

$T_{sent.}$: The number of total sentences

$C_{sent.}$: The number of successfully analyzed sentences

syntactically correct sentences to all the sentences of the original document, (analysis ratio), i.e

2.2 The Conversion and Generation Magnitude Scale (CGMS)

This scale shows the readability of translated text. In order to use this scale, you have to gather several individuals, for a panel, who must be native speakers of the target language and present them with the sets of translated documents. Only the matching sentences from the target systems that are syntactically correct are used for this evaluation.

Then, the panel judges which is a good sentence between a pair of sentences from among the three sets of the said documents, using a one-to-one comparison method [6]. Then, each evaluation sentence is ranked by each panel member. These results are then statistically calculated by the one-to-one comparison method which has been normalized on a 0 to 10 magnitude scale. This is a conversion and generation magnitude scale (described as CGMS henceforth) that can be produced using the following formula

$$M_G = 10 \cdot \frac{1}{l} \cdot \frac{1}{m} \cdot \frac{1}{(n-1)} \sum_{i=1}^{l} \sum_{j=1}^{m} (N_{ij} - 1) \qquad \text{...} (2)$$

l . Amount of evaluation sentences

m : Amount of panel members

n : Amount of evaluated systems

N_{ij} : Order number given by panel j for each translated sentence i on the system

2.3 The Translation Ability Index

Finally, the translation ability index can be produced by multiplying the analysis ratio by the CGMS as defined by the following,

$$M_T = M_A \times (M_G + 1) \qquad \text{.........} (3)$$

3 The Evaluation Experiment

We have two different calculations which arise from our evaluation method. One calculation produces the relative position for the MT system being evaluated against the ideal MT system. The other calculation produces the relative positions between the two MT systems under evaluation. The later type can actually examine several systems at once, but this was not done in this work. The systems which were evaluated with this method were English to Japanese machine translation systems.

3.1 The Sentences for the Evaluation Criteria

We used sentences from manuals for the evaluation criteria. Five domains were used. They are: airplane maintenance, the operation of a machine tool, software installation, printer installation, and network system construction. 30 sentences were selected at random from each domain area making a total of 150 sentences.

NB: In order to measure the CGMS. Sentences were chosen such that they were both correctly translated by the two MT systems, SA and SB, according to the Source-Text Analysis ratio.

The rough distribution of the used sentences that are from group (1-10 words) to (41-50words) is shown below.

Length of sentence	1 ~ 10	~20	~30	~40	~ 50 words	Total
Quantity of sentences	35	80	26	7	2	150 sen-tences

3.2 The Panel as a Representative User

We chose five members for the panel who had no experience in developing machine translation systems, but were our colleagues in the R&D division.

3.3 The Evaluator

The evaluators were four members chosen from our machine translation development team. They judged not only the goodness of sentence quality to obtain the Source-Text Analysis ratio, but also explained the judging criteria of sentences and the procedure of the one-to-one comparison method to the panel that produced the CGMS.

4 The Result of the Evaluation

4.1 The Source-Text Analysis Ratio

The results of this ratio for the two MT systems, S_A and S_B, which are sold in the market are shown below.

$$M_A(S_A) = 55.3\ (\%)$$
$$M_A(S_B) = 52.0\ (\%)$$

These systems have only about half the correct analysis ratio.

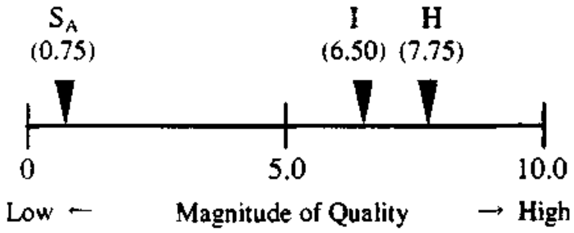
And, you can understand from definition of sets of texts, obviously results for the human translated and Ideal MT system translated sentences are,

$$M_A(H) = 100(\%)$$
$$M_A(I) = 100(\%)$$

4.2 The CGMS

There were few matching correct sentences in MT systems A and B in the domain of airplane maintenance. Therefore, this domain was omitted. In the end, 20 sentences were used, making a total of 5 sentences for each of the four remaining domain areas.

This is the result which was achieved using the method described above for MT system A.



i.e.:

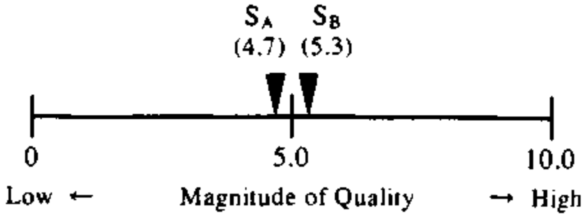
$M_G(S_A) = 0.75$

$M_G(I) = 6.50$

The results are very poor in that the $M_G(S_A)$ is lower than one magnitude, and the $M_G(S_A)$ is quite different from $M_G(I)$, the ideal MT system.

We try another evaluation method which is modified the above it. Using this modified evaluation technique it is very simple to evaluate the output of several MT systems without the necessity of preparing output from an Ideal MT system and Humans. Each M_G is also gotten from formula (2): The result of this method shows the ranking of ability of each MT system together on graph.

Next is the result of the second type of evaluation which compares the two systems, A with B.



You could get same result in case of there would be only two systems by this method; the evaluator shows panels two output and let them judge whether which one could be better. Better one is added “1”, another one is added “0”, getting the average making that 10 times, you could get same result.

We were not able to evaluate system B with the first method shown above, because there was not enough time. We, however, estimated the Mg of system B indirectly by using the data points of system A shown above from both methods. For example, if the phenomena of system B makes a proportion with the result of system A. $M_G(S_B)$=0.7725, because the result of the second type shows a difference of ± 3% between both systems. At this point, $M_G(I)$ is 6.4775, and since the Human’s Magnitude does not interact with this estimate, $M_G(H)$ remains 7.75.

4.3 The Translation Ability Index

Although, both systems, A and B, are so very poor as to be worthless, the index which is calculated for both systems is effective for considering this method.

For the case of the S_A system,

$M_T(S_A) = 55.3 \text{ x } (0.75 + 1) = 96.8$

For the case of the S_B system,

$M_T(S_B) = 52.0 \text{ x } (0.78 + 1) = 92.6$

Their performance is about the same.

4.4 Entering into Evaluation Formula for Getting Translation Ability Index

In general, for evaluation by panels, the one-to-one comparison method requires a rating of 5 ranks (ranging from almost even, a little different to obviously different) on the quality between X and Y. From this viewpoint, the subjective reactions of many people have a certain distribution which this method is able to calculate. This is a good method when you judge quality using the five human senses, smell, touch, taste, hearing, and sight. However, for evaluating the best criteria for industrial products or physical objects, then, for panels, it is better to use a judgment method based on 1 or 0 choice (e.g. better/no).

We decided to take the Translation Ability Index from formula (3), after considering the two formulas, i.e.

$$M_T' = M_A \times M_G \quad \ldots\ldots\ldots\ldots\ldots (3')$$

$$M_T'' = M_A + M_G \quad \ldots\ldots\ldots\ldots\ldots (3'')$$

The M_A and the M_G have different dimensions, but the value of the system is shown generally by multiplying components of each factor. So, formula (3') is good, and formula (3") is not. However, using formula (3') if the M_G is 0, the M_T is 0, too. This is not good because the system has its own method for handling sentence analysis, when the value of CGMS is 0. Therefore formula (3) is chosen.

5 Conclusion and Further Research

I have shown an evaluation method for the quality of a translated text, which is a one dimensional index called the Translation Ability Index. The purpose of this index is to create a method for describing the CGMS with a number, like the Source-Text Analysis Ratio. I have also presented an evaluation index or method that can be used to discuss the common domain shared by both system engineers and users.

I believe that a very objective index was created from this experiment. Although there are problems in choosing which category of documents to use, how much text, and so on, this method has a potential to be like a TOEIC or TOEFL score which are used to evaluate the ability of English usage by foreigners.

In the Japanese-English or English-Japanese translation systems, the indices which are obtained from the above numbers are shown not to be of practical use. Therefore, as a result of this work it is expected from now on that the translation ability index will be over 300 because the analysis ratio is over 70%, and the conversion and generation index will be over 3.3 (1/3 full of index) on the feast type of evaluation method. Otherwise, I hope that there will be a breakthrough in the technology which has a scale of about 5.0 for the CGMS on the second type of evaluation method to compare for Ideal MT system.

Practically, in this work a method has been produced which is more understandable and makes it easier for users to select a viable MT system.

Acknowledgments

I would like to thank Ms. Masumi Narita for her leadership in the evaluation project and the members of the MT system development team. I would also like to thank the staff of the Research & Development Group for providing the environment for our experimentation.

References

[1] ALPAC. (1966). “An Experiment in Evaluating the Quality of Translations.” LANGUAGE AND MACHINES — COMPUTERS IN TRANSLATION AND LINGUISTICS — Appendix 10, 67-75.

[2] ETL and Kyoto University. (1986). “Study about Japanese-English MT-system for Documents of Science & Technology —Report about Development of Language Processing System —.” pp.483-539.

[3] Narita, H. (1988). “Evaluation of Machine Translation Systems with Respect to the Capacity for Processing Structures.” IPSJ. NL_69-1. pp. 1-9.

[4] Takayama, T. Itoh, E. Yagisawa, Y. Mogi, K. and Nomura.H. (1993). “JEIDA’s Proposed Method for Evaluating Machine Translation (End user System Selection) — A System Questionnaire for End Users —.” IPSJ. NL_96-10. pp.73-80.

[5] Isahara, H. Shinnou, H. Yamabana, K. Moriguchi, M. and Nomura. H. (1993). “JEIDA’s Proposed Method for Evaluation Machine Translation (Translation Quality) — A Proposed Standard Method and Corpus —.” IPSJ. NL_96-10. pp.81-88.

[6] Nishizato, S. (1982). “Numerization of Qualitative Data — Dual Criterion Method and Practices —.” Tokyo: Asakura-Syoten. pp.162-171.

Appendix:

Example of using documents in this examination

Example:

Original: If there is insufficient coolant flow to the spindle, the spindle temperature will rise activating the safety interlock to prevent the saw from operating.

H：もしスピンドルへの冷却水量が不足すれば、スピンドルの温度が上昇し、保護回路が働き切断動作を停止します。

I：スピンドルへの冷却水が不十分ならば、スピンドルの温度が上昇し、保護回路が作動し、切断動作が作動するのを妨げる。

S_A：不十分な冷却水流れがスピンドルにあるならば、ノコギリが作動するのを妨げるために保護回路を活発化してスピンドル温度を増す。

S_B：スピンドルへの不十分な冷却水があるならば、スピンドル温度は、上昇するであろう（ノコギリが作動することを妨げるために、保護回路を活動的にしている）。