Quantitative Evaluation of Machine Translation using Two-way MT

Shoichi YOKOYAMA (Yamagata Univ.). Akira KUMANO (Toshiba).
Masaki MATSUDAIRA (Oki), Yoshiko SHIROKIZAWA (JST).
Mutsumi KAWAGOE (Matsushita). Shuji KODAMA (Fujitsu),
Hideki KASHIOKA (ATR), Terumasa EHARA (NHK),
Shinichiro MIYAZAWA (Shumei Univ.), and Yasuo NAKAJIMA (AAMT)

(Faculty of Engineering. Yamagata University.
4-3-16. Jonan, Yonezawa, Yamagata 992-8510. Japan
Tel. +81-238-26-3336. FAX +81-238-26-2082,
e-mail: yokoyama®emtsun.yz.yamagata-u.ac.jp)

Abstract

One of the most important issues in the field of machine translation is evaluation of the translated sentences. This paper proposes a quantitative method of evaluation for machine translation systems. The method is as follows. First, an example sentence in Japanese is machine translated into English using several Japanese-English machine translation systems. Second, the output English sentences are machine translated into Japanese using several English-Japanese machine translation systems (different from the Japanese-English machine translation systems). Then, each output Japanese sentence is compared with the original Japanese sentence in terms of word identification, correctness of the modification, syntactic dependency, and parataxes. An average score is calculated, and this becomes the total evaluation of the machine translation of the sentence.

From this two-way machine translation and the calculation of the score, we can quantitatively evaluate the English machine translation.

For the present study, we selected 100 Japanese sentences from the abstracts of scientific articles. Each of these sentences has an English translation which was performed by a human. Approximately half of these sentences are evaluated and the results are given. In addition, a comparison of human and machine translations is also performed and the trade-off between the two methods of translation is discussed.

Keywords: quantitative evaluation, two-way machine translation, word correspondence, modification, comparison of score.

1 Introduction

One of the most important issues in the field of machine translation is evaluation of the translated sentences. Some of the so-called second generation machine translation systems have adopted and improved on the evaluation method introduced in the ALPAC report. For example, the Mu project, which has developed a Japanese-English machine translation system in Japan, proposed a method of evaluation and defined five degrees of understandability and seven degrees of faithfulness[2]. The degree of understandability is evaluated by English native speakers and the degree of faithfulness is evaluated by bilingual translators. In their evaluation, the Mu project found that some examples were translated with good understandability. but with bad faithfulness, while other examples were translated with good faithfulness, but with low understandability.

In the Mu project's evaluation, the highest score of understandability. degree 1. means that "the meaning of sentences is totally clear and undoubted. Grammar, words, and styles are suitable, and therefore no correction is required." On the other hand, the lowest score of understandability. degree 5. means that "no one can understand the sentences. Even if one thinks over and discusses them, one cannot grasp their meaning." The highest score of faithfulness, degree 0. means that "the structure of input sentences is faithfully reproduced by the output sentences. English native speakers can understand them, and no or only a few corrections are needed[2]." However, the Mu project's method of evaluation is performed by humans, and therefore the evaluation is unstable. In addition, this method of evaluation is not quantitative, but qualitative.

This paper proposes a quantitative method of evaluation for machine translation systems. The proposed method is as follows. First, an example sen-

tence in Japanese is machine translated into English using several Japanese-English machine translation systems. Second, the output English sentences are machine translated into Japanese using several English-Japanese machine translation systems (different from the Japanese-English machine translation systems). Then, each output Japanese sentence is compared with the original Japanese sentence in terms of word identification, correctness of the modification, syntactic dependency, and parataxes. An average score is calculated, and this becomes the total evaluation of the machine translation of the sentence.

The proposed method of two-way machine translation and calculation of a score provides a quantitative evaluation and allows Japanese native speakers to evaluate machine translated English sentences.

For the present study, we selected 100 Japanese sentences from the abstracts of scientific articles. Each of these sentences has an English translation which was performed by a human. Using the proposed method, approximately half of these sentences are evaluated, and the results are given. In addition, we also compare human and machine translations. That is, human-machine translation is compared with machine-machine translation, and the trade-off of both translation methods is discussed.

2 Background

The present study is part of the ongoing research being conducted by the Network translation research group. The Network translation research group was established in 1997 after the reorganization of the old System evaluation work group which was organized under the AAMT (Asian-Pacific Association for Machine Translation). The old work group analyzed and studied the evaluation of sentences in Japanese source language for machine translation (MT)[5, 6], Their research showed that, in the process of sentence evaluation, the symbols included in sentences prevent correct morphological and/or syntax analysis.

In the present study, we consider the evaluation of machine translation systems using natural language processing concepts[3, 4]. This is because in a machine translation system, almost all of the various processing is based on natural language processing concepts.

The purpose of the present study is as follows:

- To 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.
- To 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 these issues, to establish new concepts for machine translation.
- To collect sentences difficult to machine-translate and make these fundamental test-beds for examination of the capabilities of machine translation
- To analyze the transfer part of the machine translation process by comparing bilingual lexicons, and improve the transfer lexicons.
- To create alternative sentences to the original ones, and also use these sentences as candidates for testing the capabilities of machine translation.

The syntax and semantics standards of sentences for the purpose of system developers [1] have already been established by the JEIDA (Japan Electronic Industry Development Association). Various sentences difficult to translate have been collected and analyzed [5,6] by the System evaluation work group. The purpose of this was so that the user could avoid inputting the sentences difficult to machine translate, and to establish guidelines for the user such as, "proper nouns frequently used should be registered in the dictionary." or "special adjective suffixes should not be used if possible. However, these are only brief and qualitative evaluation standards or advices.

As a result of rapid development in network communication, the need for machine translation and/or machine assisted translation is increasing. Therefore, a quantitative and subjective method of evaluating machine translation systems is required.

3 Procedure and Method of Evaluation

Being Japanese native speakers, it is difficult for us to evaluate the quality of English of Japanese-English machine translation. Therefore, in the present study, Japanese sentences are first translated to English, and then English to Japanese in order for to us to compare the output Japanese sentence with the original sentence.

Fig. 1 shows a flow diagram which summarizes the procedure. In order to evaluate sentences quantitatively and objectively, the procedure is as follows.

1. 100 Japanese sentences are randomly selected from the abstracts of articles in computer science. Each of these sentences is translated

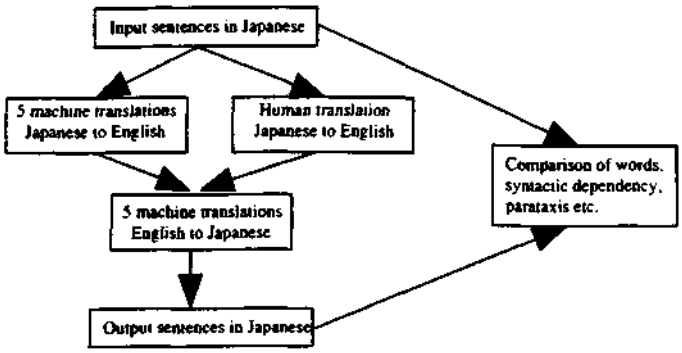

Figure 1: Evaluation procedure for machine translation

into English by a human. These English sentences are used for reference.

2. The Japanese sentences are machine translated into English using five different commercial systems without pre-editing. Normally, a few sentences cannot be translated or are only partly translated because of the performance of the system. The cases in which no translation results are obtained are ignored. However, results with only partial translation are included.
3. The output English sentences are machine-translated back into Japanese using five commercial systems. These five English to Japanese machine translation systems are basically independent from the five Japanese to English systems described above. The human translation is also machine-translated for comparison.
4. Each output Japanese sentence is compared with the original Japanese sentence, and an evaluation score is calculated based on the criteria given below.

Approximately 3.000 sentences were obtained using the above procedure. This is because six English sentences including the human translation are output from one Japanese sentence, and from these 30 resulting Japanese sentences are obtained finally. Currently, approximately half of these sentences have been evaluated. The practical evaluation items are as follows.

(A) Correspondence of words: counts the following number of words which exist in the output sentence corresponding to the original sentence:

1. the number of complete corresponding words (A0).
2. the number of partially corresponding words (A1), and
3. the number of homonyms (A2).

(B) Correspondence of modification: counts the number of correct connections between the modifier and the modified constituents.

(C) Correspondence of parataxis: counts the number of parataxes.

A score is calculated from the above three items. For the correspondence of words, complete correspondence (A0) is counted as 1.0. and partial correspondence (A1) and homonym (A2) are counted as 0.5. respectively. These three scores are summed, and then a score A is obtained. The A. B. and C scores are compared with the number of words, modification, and parataxes of the original sentence, respectively.

Finally, the average scores of machine-machine translation and human-machine translation are calculated, and then compared.

4 Evaluation Examples

An example is shown in Fig. 2. In the Figure. "#" followed by a Japanese sentence is the original sentence from an abstract of a scientific article, and "#" followed by English sentence is the corresponding human translation. The sentence contained in parentheses after the Japanese sentence represents a romanization (Japanese method) of the Japanese. "%" followed by an English sentence represents an example of the machine translation. Each English sentence is followed by Japanese sentences which were machine-translated by the commercial systems (in the Figure, only a selection of the results are shown).

In the Figure. (A). (B). and (C) represent the evaluation items described in the previous section. (A0)

means complete correspondence, (A1) means partial correspondence, and (A2) means homonyms, as mentioned above.

In the example shown in the Figure, the original Japanese sentence consists of 19 main words, and includes 8 modification relations and 1 parataxis. The sentence starting with "#By comparing ..." is the human translation of the original Japanese sentence. The following two Japanese sentences are examples of the results of the English-Japanese machine translation systems with which the scores of A0 - C1 are calculated. In the first Japanese sentence, the score for the complete correspondence of words is 12, partial correspondence is 2. and homonyms is 0. The score A is 13, and the number of modification relations is 8. That is. all relations included in the original sentence also exist in this sentence. The number of parataxes is 1. These scores show the results of the human-machine translation.

The sentence starting with "%By comparing ..." is a result of a Japanese-English machine translation system. The following two Japanese sentences are examples of the results of the English-Japanese machine translation systems. The method of score calculation used is the same for both. These scores show the results of machine-machine translation.

#ＩＴＵ通信評価基準に基づき各国のディジタル通信量を比較することで，各国のマルチメディアの普及度と経済レベルの相関関係がよく分かる。

(ITU tûsin hyôka kizyunni motoduki kakkokuno dezitaru tûsinryôwo hikakusurukotode, kakkokuno marutimedeano hukyûdoto keizai reberuno sôkan kankeiga yoku wakaru.)

(A) 19words, (B) 8, (C) 1

#By comparing digital communication quantity of each country based on the ITU communication criterions for evaluations, the correlation between prevalence and economy levels of multi-media in each country is well proven.

評価用の ITU 通信 criterions に基づいた、各国のディジタル通信量の比較によって、各国の多重媒体の普及と経済のレベル間の相関は、上手に証明されている。

(Hyôkayôno ITU tûsin criterionsni motoduita, kakkokuno dezitaru tûsinryôno hikakuni yotte, kakkokuno tazyû baitaino hukyûto keizaino reberukanno sôkanwa, zyôzuni syômei sareteiru.)

(A0) 12, (A1) 2, (A2) 0, (B) 8, (C) 1

評価のための ITU 通信判定基準に基づく個々の国のデジタル通信量を比較することによって、個々の国のマルチメディアの普及と経済レベルの相関がよく証明される。

(Hyôkayôno tameno ITU tûsin hantei kizyunni motoduku kokono kunino dezitaru tûsinryôwo hikakusurukotoni yotte, kokono kunino maruti medeano hukyûto keizaino reberuno sôkanga, yoku syômeisareru.)

(A0) 13, (A1) 1, (A2) 2, (B) 7, (C) 1

%By comparing digital communicative quantity of each country based on ITU communicative evaluation standards, a correlation of a widespread degree of MultiMedia of each country and an economic level is found well.

ITU のコミュニケーションの評価規格に基づくデジタルコミュニケーションの量の各国を比較することによって、各国の MultiMedia の広範囲の度と経済レベルの相関関係はよく見つけられる.

(ITU komyunikêsyonno hyôka kikakuni motoduki dezitaru komyunikêsyonno ryôno kakkokuwo hikakusurukotoni yotte, kakkokuno MultiMediano kôhan'ino doto keizai reberuno sôkan kankeiwa yoku mitukerareru.)

(A0) 13, (A1) 0, (A2) 0, (B) 5, (C) 1

ITU 通信評価標準に基づくデジタルの通信量の個々の国を比較することによって、個々の国と経済水準の広範囲に及んだ程度のマルチメディアの相関がよいのが発見される。

(ITU tûsin hyôka kizyunni motoduki dezitaruno tûsinryôno kokono kuniwo hikakusurukotoni yotte, kokono kunito keizai suizyunno kôhan'ini oyonda teidono marutimedeano sôkanga yoinoga hakkensareru.)

(A0) 12, (A1) 0, (A2) 3, (B) 2, (C) 0

Fig. 2 Example sentences and score calculation

Based on the results, the calculated scores for the example shown in Fig. 2 are shown in Table 1. In the Table, the row labeled "original" contains the number of words (A), the number of modification relations (B), and the number of parataxes (C) for the original sentence. These are also shown at the top of Fig. 2.

The row labeled "human" contains the average scores of the human-machine translation. The score (A) is first calculated using the method described above, and then the five output scores are averaged. The scores (B) and (C) are the average of the number of modification relations and the average of the number of parataxes for the results of the five systems, respectively.

The row labeled "machine" contains the average scores of the machine-machine translation. In the Table, the scores shown are the average of 25 results. The method used to calculate the score is the same as the human-machine translation.

Table 1 shows that in the translation of the Japanese sentence shown in Fig. 2, the score of human-machine translation is higher than the score of machine-machine

translation. For the correspondence of words, approximately 69% of words correspond in human-machine translation, whereas approximately 59% of words correspond in machine-machine translation.

Table 1 The scores of the example sentence shown in Fig. 2

	(A)	(B)	(C)
original	19	8	1
human	13.1	3.6	0.8
machine	11.3	3.7	0.24

Table 2 shows the comparison of human-machine and machine-machine translation. Of the 100 Japanese sentences 40 were evaluated, of which 28 were sentence style, and 7 were noun phrases such as article titles. Only 5 sentences were compared with the machine-machine translation. In Table 2. only the comparison for score (A) is shown.

In the Table, "mt > noun" shows in how many sentences the score of machine-machine translation is greater than the score of human-machine translation. The row labeled "human > 50%"' means that the score of human-machine translation is greater than 50% of the number of words in the original sentence. That is, more than half of the words correspond to the original words. Similarly, for the row labeled "mt > 50%".

Table 2 Comparison of human and machine translation.

Sentences 28			
mt > human	9	<	19
human > 50%	15	< 50%	13
mt > 50%	13	< 50%	15
Noun phrases 7			
mt > human	1	<	6
human > 50%	3	< 50%	4
mt > 50%	2	< 50%	5
Mt evaluation only 5			
> 50%	1	< 50%	4

As shown in the Table, the score of human-machine translation is higher overall than the score of machine-machine translation both for sentences and noun phrases. However, for the cases in which the machine-machine translation is higher, human Japanese-English translation is a little problematic, or it includes new information concerning the context. For approximately half of the sentences the correspondence rate is more than 50%, and this rate is comparable for human and machine translations.

As described above, in some cases, the score of the human-machine translation is worse than the score of machine-machine translation. Fig. 3 shows a typical example using our current criteria. The meaning of the symbols used in Fig. 3 is the same as in Fig. 2.

#すなわち，通信条件ならびにNCのパラメータ設定法，接続ケーブル，割り込み処理などについて解説した。

(Sunawati, tûsin zyôken narabini NCno paramêta setteihô, setuzoku kêburu, warikomi syori nadonituite kaisetusita.)

(A) 15words, (B) 1, (C) 1

#This paper explains conditions of communication and a method for setting NC parameters, connecting cables, interrupt handing, etc.

この論文はコミュニケーションの状態とケーブル，中断の手渡すことなどを接続してNCパラメタを設定するための方法を説明する.

(Kono ronbunwa komynikêsyonno zyôtaito kêburu, tyûdanno tewatasukoto nadowo setuzokusite NC parametawo setteisuru tameno hôhôwo setumeisuru.)

(A0) 6, (A1) 1, (A2) 1, (B) 1, (C) 0

この論文は、コミュニケーションの条件<状態>と設定NCパラメーター、接続ケーブル、中断<割込み>扱い、その他のための方式を説明する

(Kono ronbunwa komynikêsyonno zyôken <zyôtai>to setteı NC paramêtâ, setuzoku kêburu, tyûdan <warikomi> atukai, sonotano tameno hôsikiwo setumeisuru.)

(A0) 7, (A1) 0, (A2) 1, (B) 1, (C) 1

%That is to say, the following were explained : Communication conditions and parameters setting method of NC, connection cables, interrupt handling, etc..

すなわち、下記は説明された: NC、接続ケーブル、割込み取り扱いなどの通信条件、およびパラメーター・セッティング方法。

(Sunawati, kakiwa setumeisareta: NC, setuzoku kêburu, warikomi toriatukai nadono tûsin zyôken, oyobi paramêtâ settingu hôhô.)

(A0) 8, (A1) 1, (A2) 0, (B) 1, (C) 1

すなわち、以下は説明された：NC、接続ケーブル、割込み操作などの方法を設定している通信条件とパラメータ。

(Sunawati, ikawa setumeisareta: NC, setuzoku kêburu, warikomi sôsa nadono hôhôwo setteisiteiru tûsin zyôkento paramêta.)

(A0) 10, (A1) 1, (A2) 1, (B) 1, (C) 1

Fig. 3 Comparison of human and machine translations

As shown in this Figure, the human Japanese-English translation adds new information that the original Japanese sentence does not contain. In the exam-

ple in the Figure. “This paper explains ...” is considered to be typical starting sentence in a paper, but the original Japanese sentence does not include this.

The score of the examples shown in Fig. 3 is shown in Table 3. The labelling and method of score calculation is the same as in Table 1. As shown in Table 3. the score of the machine translation is slightly better than the score of the human translation for all categories.

As shown in Table 3, approximately half of the words correspond to the original for both human-machine translation and machine-machine translation. The modification score is equal for both methods of translation. The score of the parataxes is much lower in the human-machine translation because new information was added to the sentence such as “this paper explains ...”.

Table 3 The scores of the examples shown in Fig. 3

	(A)	(B)	(C)
original	15	1	1
human	7.5	0.6	0.2
machine	7.98	0.6	0.64

5 Concluding Remarks

The present study has shown that quantitative and objective evaluation of machine translation is possible in terms of the correspondence of words, modifiers, and parataxis.

The proposed method of evaluation clearly shows the difference between human and machine translation. That is, when a human translates Japanese sentences into English, (s)he considers the context and the flow of the topics. (S)He may add new information and/or words, or conversely (s)he may remove part of sentences and/or words. Pronominalization and zero anaphora are also used in this case.

"Intelligible" human translation is sometimes difficult to machine translate, and human translations which take into account the context produce relatively low scores using the proposed criteria.

Using the proposed two-way method of machine translation, where Japanese sentences are machine translated into English, and then translated back into Japanese, it was found that, in some cases, many words are recovered.

Using various translation systems reduces the dispersion of the individual systems and, we believe, improves the quality of the translation.

Currently, procedures such as word derivation and deciding corresponding words are performed by humans. However in the near future, it will be possible for these procedures to be performed automatically. Already, word extraction can be automatically performed using morphological analysis programs. However, modification requires precise syntax and semantic analysis, which makes automatization difficult. If the corpus with added tags could be utilized, an automatic method would be possible.

Using the proposed method, compensation of human and machine translation will be possible. That is, by using both human and machine translation it is possible to raise the quality of translation.

Even if the score obtained using the proposed method is good, this does not necessarily mean that the translation result is good since the evaluation criteria are limited in order to simplify the evaluation. In the near future, we aim to complete the total analysis (including items (B) and (C)), *as* well as compare our method with traditional human evaluation.

References

[1] Isahara H. et al. (1996). “Technical Evaluation of MT Systems from the Developer’s Point of View: Exploiting Test-Sets for Quality Evaluation” (in Japanese), Journal of Natural Language Processing. Vol. 3, No. 3 pp.83-102.

[2] Makoto Nagao (1985). “Evaluation of the Quality of Machine Translation Sentences and the Control of Language” (in Japanese). Information Processing, Vol. 26. No. 10, Information Processing Society of Japan (IPSJ) pp.1197-1202.

[3] Shoichi Yokoyama (1992). “Toward a Systematic Evaluation of Machine Translation: from the Viewpoint of Natural Language Processing”, Proc. of International Symposium on Natural Language Understanding and AI as a Part of International Symposia on Information Sciences (ISKIT'92) pp.102 106.

[4] Shoichi Yokoyama (1993). “Evaluation Method of Machine Translation: From the Viewpoint of Natural Language Processing”, Proc. of MT Summit IV pp.215-217.

5] Yokoyama S. et al. (1994). “Collection and Classification of Sentences Difficult to Machine-Translate” (in Japanese), Information Processing Society of Japan (IPSJ). SIG-NLP, NL101-5

[6] Yokoyama S. et al. (1994). “Machine Translation and Evaluation of Japanese Sentences Difficult to Translate” (in Japanese), ibid., NL101-6