An Automatic Evaluation Method for Machine Translation using Two-way MT

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

Faculty of Engineering, Yamagata University,
4-3-16, Jonan, Yonezawa, Yamagata 992-8510, Japan,
yokoyama@emtsun.yz.yamagata-u.ac.jp

Abstract

Evaluation of machine translation is one of the most important issues in this field. We have already proposed a quantitative evaluation of machine translation system. The method was roughly that an example sentence in Japanese is machine translated into English, and then into Japanese using several systems, and that the comparison of output Japanese sentences with the original Japanese sentence is done for the word identification, the correctness of the modification, the syntactic dependency, and the parataxis. By calculating the score, we could quantitatively evaluate the English machine translation.

However, the extraction of word identification etc. was done by human, and the fact affects the correctness of evaluation. In order to solve this problem, we developed an automatic evaluation system. We report the detail of the system in this paper..

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

1. Introduction

Evaluation of machine translation is one of the most important issues in this field. Some of the machine translation systems adopted the human scoring evaluation like the evaluation method of the ALPAC report. This type of evaluation is unstable, and is not quantitative, but qualitative.

We proposed a quantitative evaluation of machine translation system (Yokoyama, 1999). The method was as follows: first, an example sentence in Japanese is machine translated into English using several Japanese-English machine translation systems. Second, output English sentences are machine translated into Japanese using several English-Japanese machine translation systems (different from Japanese-English machine translation systems). Then, the comparison of output Japanese sentences with the original Japanese sentence is done for the word identification, the correctness of the modification, the syntactic dependency, and the parataxis. The average score is calculated based on the proposed method, and it becomes the total evaluation of machine translation for the sentence.

From this two-way machine translation and the calculation of the score, we can quantitatively evaluate the English machine translation in which it is difficult to evaluate them for Japanese native speakers.

However, most of extraction and comparison processes are performed by human, and the stableness and the speed are problematic.

In this paper, we propose an automatized process of the evaluation, and show that the result is the same as by human. In addition, we calculate the score of each Japanese-English machine translation system. The result directly shows the performance of the system. We also try to calculate the score of each English-Japanese machine translation system. The result shows that some system makes a good effort even if some Japanese-English translation systems output bad results.

Practically, we selected 100 Japanese sentences from abstracts of scientific articles. Each of these sentences has a human English translation. The result is shown below. The advantage of this method is that the automatic quantitative and objective evaluation is possible. The trade-off and evaluation of both translation methods is also discussed.

2. Background

Network translation research group which follows the system evaluation working group under AAMT (Asian-Pacific Association for Machine Translation) was established in 1997. Under the old working group, we have collected and analyzed the sentence difficult to translate for users (Yokoyama, 1994a, 1994b). The study had the purpose that the user avoided to input the sentences difficult to machine translate, and that the guidelines for the user would be established. However, there are only brief and qualitative evaluation standards or advices such that "proper nouns frequently used should be registered in the dictionary," or "special adjective suffixes should not be used if possible. "

Recently, the network communication develops in the rapid, and to the huge stage, and the necessity of machine translation and/or machine assisted translation increases. The evaluation must be both quantitative and subjective. We have evaluated some machine translation systems on the network (so called "netsurf" translation systems) from the viewpoints of various in- and output forms such as button operations, marquees, and so on (Miyazawa, 1999).

We have also evaluated the machine translation system quantitatively. The issue of this system is to automatize entire process.

Recently we can use free morpheme analysis system like "Chasen" , "Juman" , and others, and also utilize free parsing system like "KNP" . We use "Juman" and "KNP" for extracting morphological and syntactic structures for Japanese sentences. We utilize classification number from "Bunrui Goi Hyou" (Classification of vocabulary in Japanese) (NLRI, 1964) for semantic analysis. Their usage becomes the establishment of totally automatized system for the evaluation of machine translation system.

3. Procedure and Method of Evaluation

3.1. Summary of Procedure

As for Japanese-English machine translation, it is difficult for us to evaluate the quality of English because we are not native speakers of English. Conversely, as native speakers, we can easily evaluate Japanese sentences.
So, the procedure of evaluation is as follows:

1. Original Japanese sentence is read in.
2. Output Japanese sentence, which is the output of Japanese-English and English-Japanese machine translation, is read in.
3. The analysis for both sentences is performed using "KNP" which calls "Juman". Morphemes and syntax are analyzed.
4. The comparison for each word is done.
5. The contrasting for modification is done.
6. The result is output and the score is calculated.

Fig.1 shows the summary of the procedure. In order to evaluate sentences quantitatively and objectively, the procedure goes as follows:

(a) 100 Japanese sentences are randomly selected from abstracts of articles in computer science. Each of these sentences is translated into English by human. We use these English sentences as reference.
(b) The Japanese sentences are machine-translated into English using 5 different commercial systems without pre-editing. A few sentences cannot be translated or are only partly translate because of the performance of a system. We ignore the cases in which no translation results is obtained, but use the results with only partial translation.
(c) The output English sentences are machine-translated back into Japanese using 4 commercial systems. These 4 English to Japanese machine translation systems are basically independent to 5 Japanese to English systems described above. Human translation is also machine-translated for comparison.
(d) The output Japanese sentences are compared with the original Japanese sentence, and the evaluation score is calculated based on the criteria mentioned below.

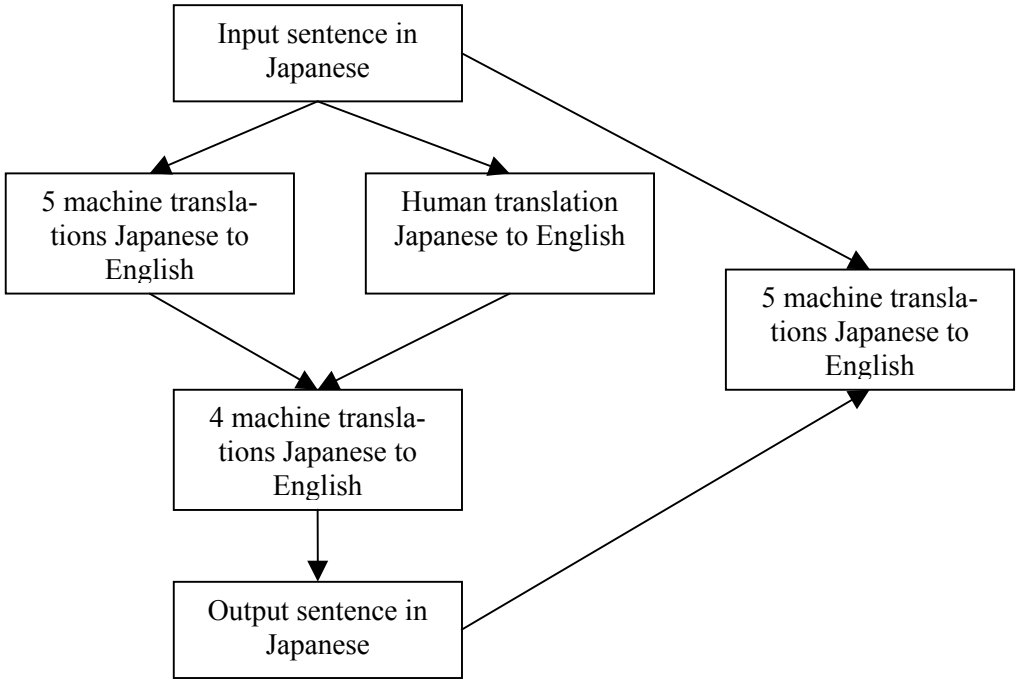

Fig.1 Evaluation procedure for machine translation

About 2,400 sentences are obtained by the above procedure because 6 English sentences including a human translation are output from one Japanese sentence, and 24 resulting Japanese sentences are obtained finally.

3.2. Correspondence of Words

The automatic process of correspondence of words between an original and an output Japanese sentence is performed as shown in Fig.2. Program is written in Perl usable for Japanese.

First, the complete correspondence of a word in an original and an output sentence is done. This matching process is very simple.

Second, the partial correspondence is performed. In this process, the longest common strings between words are calculated using the dynamic programming, and if more than 60% of the average length of two words are the same, the partial correspondence is adopted.

For example, assume that in

W_1: $\underline{C_{11}C_{12}}C_{13}C_{14}$ and

W_2: $C_{21}\underline{C_{22}C_{23}}C_{24}$,

if the strings $C_{11}C_{12}$ equal to the strings $C_{22}C_{23}$, the rate is calculated as 2*2/(4+4) = 0.5, that is 50%.

The correspondence of semantics is used "Bunrui Goi Hyou" (Classification of vocabulary in Japanese) (NLRI, 1964). It is a thesaurus in which entry word, classification number (for example, 1.100), sub-classification numbers. Now we use the classification number, and if the number is the same, that is, two words are included in the same category, the correspondence is regarded as formed. Practically, we are using the old version of the thesaurus, the matching rate is relatively low because of few vocabulary.

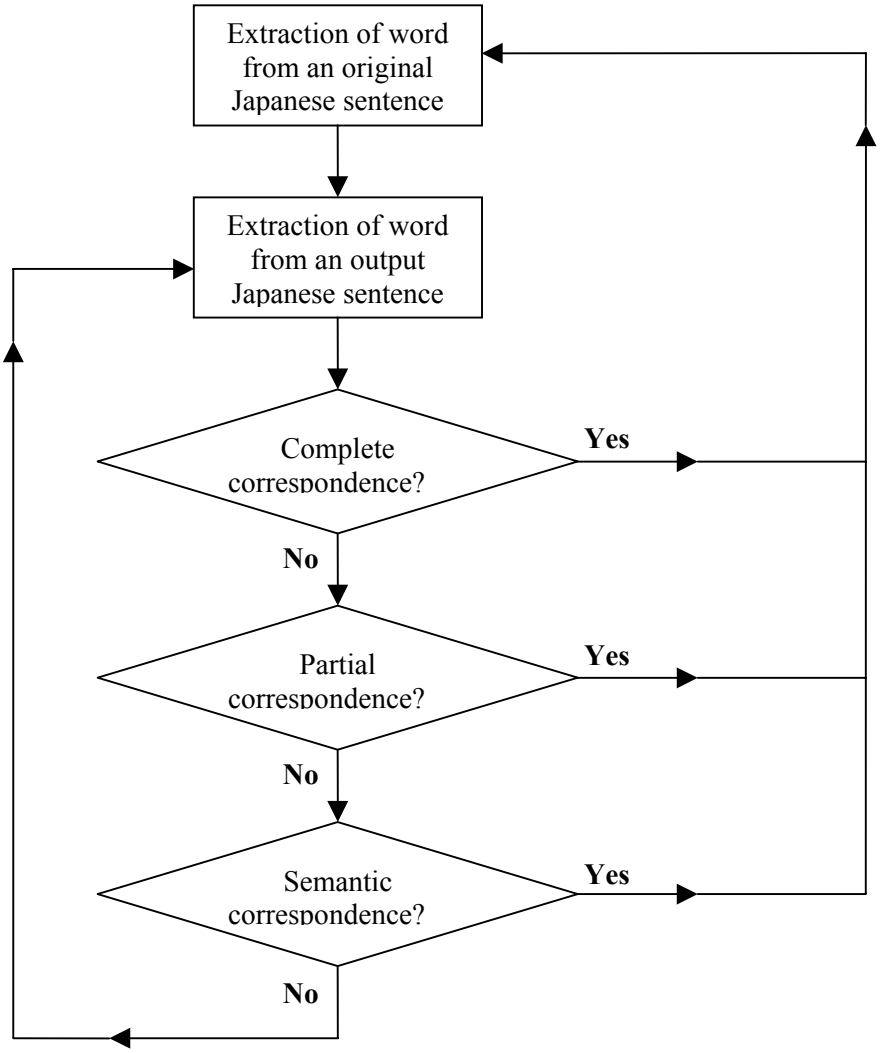

Fig.2 Procedure for correspondence of words

3.3. Correspondence of Modification

In Japanese, the relation between a modifier and a modifee is very important because the semantic relationship is found in it. We use a free parser "KNP", and the structure of each sentence is made clear. "KNP" shows the relation such as parataxis, appositions, and simple modification. We extract the correspondence of modified relations between an original and an output sentences. Now we are trying to use this parameter, so that we do not show its result in this paper.

4. Examples of Evaluation

An example is shown in Fig.3. In the Figure, "[#3]" etc. shows the number of input sentence. As shown in the Figure, the original Japanese sentence written in romanization is first shown, followed by human translation into English as a reference, which is translated considering the context of the sentence, by a machine translation into English (in this Figure, using System A), and by a output Japanese sentence translated from the English machine translated sentence. At last, evaluation of comparison is shown. "[#3]" is the same example as shown in the previous paper (Yokoyama,1999), but since the version and kind of systems are totally different one another, the simple comparison is impossible. The result is relatively good as shown in the Figure. "[#22]" obtains a relatively good result, but on the other hand, "[#61]" obtains a bad result. "[#22]" is relatively a long sentence, but since there are several repetitions of the same word and the syntactic structure is relatively simple, the good result is introduced. On the contrary, "[#61]" is a short sentence. However, since the English machine translation uses not so good corresponding words, and the Japanese translation directly translates these words, the corresponding rate becomes low.

[#3]

Japanese Input Sentence

ITU tuusin hyouka kizyunni motoduki kakkokuno dezitaru tuusinryouwo hikakusurukotode, kakkokuno marutimedeano hukyuudoto keizai reberuno soukan kankeiga yoku wakaru.
(A) 21words (B) 12

Human translation into English

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.

A machine translation into English

The correlation at diffusion and an economic level of the multimedia in each country is understood from the comparison of the amounts of digital communication in each country based on ITU communication criterion well.

Output Japanese sentence

Kakusanni okeru soukan kankeito kakkokuno maruti mediano keizai reberuwa ITU komyunikeisyon hyouka kizyunni yoku motoduku kakkokunde dezitaru tuusinno ryouno hikakukara rikaisareru.
(A) 20words (B) 13

Evaluation

(A0) 16 (A1) 0 (A2) 0 (B) 3

[#22]

Input Japanese sentence

Kenkyuu mokutekiwa, waado purosessa sousano hyouzyunka sonomowo zissisurukoto dewanaku, sousano hyouzyunkatoiu sigotowo suisinsiteiku uede hituyouninaru gizyututekina houhouwo zyunnbisuru kotoniaru.
(A) 22words (B) 16

Human translation into English

The purpose of the research is not in realizing standardization itself of word-processor operations but in preparing technical methods to be necessary in promoting standardization of operations.

A machine translation into English

The research purpose is in the preparation for the technical method needed in promoting work of standardization of not the execution of the standardization of the word processor operation but the operation.

Output Japanese sentence

Waado purosessa sousano hyouzyunkano zikkoudewanaku, sousano hyouzyunkano sigotowo sokusinsuruniioite hituyoudearu gizyututekina houhounotameno zyunbiniwa kenkyuu mokutekigaaru.
(A) 19words (B) 14

Evaluation

(A0) 16 (A1) 0 (A2) 0 (B) 7

[#61]

Input Japanese sentence

Tuusin karaoke saabisuwo VTX moude kaisi.
(A) 5words (B) 2

Human translation into English

#Starting the communication karaoke service through VTX network.

A machine translation into English

#The interactive karaoke service begins with the VTX net.

Output Japanese sentence

Taiwatekina karaoke saabisuwa VTX nettode hazimaru.
(A) 5words (B) 3

Evaluation

(A0) 2 (A1) 0 (A2) 0 (B) 0

Fig.3 Example sentences and the score calculation

In the Figure, (A) shows the number of independent words extracted from the "Juman", and (B) shows the number of modifications extracted from the "KNP". In addition, (A0) means complete correspondence, (A1) means partial correspondence, and (A2) means homonyms as mentioned above.

As shown in Fig.3, the score is automatically calculated. In Fig.3, the result is relatively good because the correspondence of words is very high.

We totalize various extracted relations described above. The result is shown in Table 1.

N	of total	words	1507			
N	of total	mod.	886			
		a	b	c	d	average
A	total w.	1659	1801	1740	1705	1726
	total m.	1089	1248	1116	1075	1132
	cor. w.	847	716	800	749	778
	par. w.	181	192	284	191	212
	cor. m.	400	492	360	258	378
B	total w.	1606	1773	1616	1653	1662
	total m.	1102	1254	1053	1099	1127
	cor. w.	741	699	705	742	722
	par. w.	170	175	272	178	199
	cor. m.	306	490	214	318	332
C	total w.	1694	1794	1725		1738
	total m.	1190	1304	1182		1225
	cor. w.	697	647	720		688
	par. w.	169	185	294		216
	cor. m.	258	190	248		232
D	total w.	1751	1851	1796	1747	1786
	total m.	1185	1331	1216	1167	1225
	cor. w.	785	704	797	773	765
	par. w.	187	162	269	172	198
	cor. m.	332	244	314	332	306
H	total w.	1639	1820	1672	1678	1702
	total m.	1165	1284	1140	1160	1187
	cor. w.	711	660	735	706	703
	par. w.	183	195	236	181	199
	cor. m.	238	434	264	228	291

Table 1 The total score of the sentences

The number of total words in original 100 Japanese sentences is 1507, and the number of total modifications found by using "KNP" is 886, as shown in Table 1. A capital letter shows a Japanese-English system, and a small letter shows an English-Japanese system. "H" means human translation into English.

"Average" shows the average of a line, that is, the average of a Japanese-English translation system. For example, the number of total words is 1659 in A-a of the output Japanese sentences in which original Japanese sentences are machine translated by the system A, and the output English sentences are translated into Japanese sentences using the system a.

As shown in Table 1, usually the number of total words and total modifications are more than the number in the original Japanese sentences.

Table 1 also shows the performance of a Japanese-English system. The performance of the system A is totally better than the system B because each and the average score of the correspondence of words and the correspondence of modifications are better in the system A than in the system B.

5. Concluding Remarks

From the viewpoint of the correspondence of words, modifiers, and parataxis, the quantitative and objective evaluation of machine translation is possible.

Automatic evaluation procedure is established by this total system because the evaluation method is performed automatically.

However, there are some issues to solve. For example, the algorithm of correspondence of words is advanced from the top of the sentence to the last, and then if there are more than one corresponding word, it is difficult to decide which word must be matched. The problem also

occurs in the matching of modification. It is also difficult to decide which phrases should be matched.

In this paper, the semantic correspondence is not so much written because the vocabulary included in "Bunrui Goi Hyou" is very few for the scientific articles. The definition of semantics should be discussed in the future.

We will continue to refine the total system, and establish automatic evaluation of machine translation.

Bibliography

The National Language Research Institute (NLRI) (ed.) (1964) "Bunrui Goi Hyou" (Classification of vocabulary in Japanese) (in Japanese). Shuei Publishing Co., Tokyo

"Juman" (free morphological analysis system), Kyoto University

"KNP" (free parsing system), Kyoto University

"Chasen" (free morphological analysis system), Nara Institute of Science and Technology

Miyazawa S. et al. (1999). Study on Evaluation of WWW MT Systems. Proceeding of Machine Translation Summit VII, 290-297.

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

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

Yokoyama S. et al. (1999). Quantitative Evaluation of Machine Translation Using Two-Way MT. Proceeding of Machine Translation Summit VII , 568--573.