Evaluation of MT Systems by TOEFL

Masaru Tomita
Faculty of Environmental Information, Keio University
School of Computer Science, Carnegie Mellon University
mt@sfc.keio.ac.jp

Masako Shirai
Sun Microsystems Computer Corporation
masako.shirai@Japan.Sun.COM

Junya Tsutsumi
Department of Computer Science, Keio University
junya@nak.math.keio.ac.jp

Miki Matsumura
Department of Environmental Information, Keio University
t90514mm@sfc.keio.ac.jp

Yuki Yoshikawa
Department of Environmental Information, Keio University
t90624yy@sfc.keio.ac.jp

Abstract

Recently, we did some comparative evaluation of the commercial English-to-Japanese MT systems by using the TOEFL, Test of English as a Foreign Language. The passages in the Reading Comprehension part of the Vocabulary and Reading Comprehension section were extracted from a TOEFL guide book [1], and were translated into Japanese by using the MT systems. Then the translated tests were taken by the examinees who are native speakers of Japanese as if they were Japanese comprehension tests. The scores of the tests would reflect the systems' abilities to convey the semantic contents of the input passages to the examinees. This paper describes what we have learned from the experiments.

1. Introduction

The MT systems have been evaluated by the MT developers, prospective users, and some technical publishers. And the most common method is the sentence-by-sentence evaluation [2] in the following manners such as that: each translated sentence is evaluated, or the post-editing process is measured. While the sentence-by-sentence evaluation can certainly give some reasonable measure of the quality of translation, we are often concerned with the quality of translated text, a group of sentences, as a *whole.*

As we elaborate in Section 2, "What is the TOEFL?," the TOEFL has a section to test the reading comprehension. Examinees are given five or six passages, each one followed by four to six questions. The better the examinee understands the passage, the more questions he or she can answer correctly, resulting a higher score.

Several reading comprehension parts from a TOEFL guide book [1] are selected, and translated by English-to-Japanese MT systems. The results are Japanese comprehension tests: each test consists of a passage in Japanese followed by the multiple choice questions in Japanese. The Japanese tests are taken by native speakers of Japanese. If the original English passages are correctly translated, the examinees would have little problem in understanding the passages, and therefore make the perfect or near perfect scores. If, on the other hand, the passages are translated poorly, the examinees would hardly understand the passages, and therefore make the low scores. Thus, these scores by the examinees, native speakers of Japanese, would reflect the quality of translation.

Section 3, "Evaluation Method," describes the details on how we evaluated the MT systems.

Section 4, "Results," describes the results of the evaluation.

Section 5, "Discussions," interprets the results, and discusses the problems.

Finally, Section 6, "Concluding Remarks," summarizes the experiments, and concludes this paper with the discussions about the issues to be addressed in our future evaluation.

2. What is the TOEFL?

The purpose of the TOEFL is to measure the examinee's ability to read and understand North American English. Many colleges and universities in the United States and Canada require their applicants to take the TOEFL, if the applicants are not native speakers of English. In addition, many other institutions, government agencies, and scholarship programs use TOEFL scores. In some countries the TOEFL is used to assess a person's knowledge of English for job purposes within that country. TOEFL can be taken 12 times a year in almost every major city world wide.

In general, a score of 600 or more is considered excellent and a score of 400 or less is weak. The highest possible score anyone can get is 677. Some universities require a score of 550 or higher for entrance.

TOEFL includes three sections: Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension. In this evaluation, we are interested only in the Vocabulary and Reading Comprehension section, especially in the Reading Comprehension part. Usually the Reading Comprehension part of the TOEFL contains five or six different passages which can be on any topic, each one followed by four to six questions. The following reading topics frequently appear in the Reading Comprehension section of the TOEFL [3].

Sciences
Arts and Literature
United States History and Government
Other Social Sciences

As shown in the above list, the TOEFL tests offer variety of topics in its Reading Comprehension section. In addition, the TOEFL passages are written in a very brief manner. They contain only a little more than the basic information that an examinee need to answer the questions. For these reasons, the TOEFL tests are suitable for the materials to measure the accuracy of translation of the passages.

The example tests are shown in "Appendix A."

3. Evaluation Method

Three commercial MT systems were evaluated in this evaluation, referred as Evaluation 1. As mentioned in Section 1, "Introduction," we have used the TOEFL tests as input texts for translation.

The evaluation was done according to the following steps:

1. The Reading Comprehension parts were extracted from four TOEFL simulation tests from a TOEFL guide book [1], and translated by the MT systems (listed as System A, B, and C in Table 1). The texts were also translated manually (listed as Human in Table 1). One unit consists of a set of three to four passages each.
2. Each of the translated unit was divided into three parts, and pasted together as shuffled so that each unit would equally consists of all system's translation.
3. The questions and multiple choices texts were translated manually.
4. The tests were taken by 60 students who are native speakers of Japanese. This gave us 20 examinees per system.
5. The tests were graded and the scores for each system were calculated. The higher the score, the better the system performed.

The translation examples are shown in "Appendix B."

4. Results

This section describes the results of Evaluation 1. As far as analyzing the results of Evaluation 1, none of the MT systems we evaluated perfectly translated the passages, but the examinees basically understood the outlines of the passages.

The following table lists the percentages of the correct answers:

Table 1: Percentages of Correct Answers (Evaluation 1)

	Test 1	Test 2	Test 3	Test 4	Average
System A	41%	57%	42%	78%	55%
System B	52%	57%	37%	78%	56%
System C	50%	60%	39%	82%	58%
Human	100%	70%	100%	90%	91%

The following table lists the estimated TOEFL scores that are converted from the raw scores resulted from the testing:

Table 2: Estimated TOEFL Scores (Evaluation 1)

MT Systems	TOEFL Scores
System A	465
System B	471
System C	477
Human	620

According to the estimated TOEFL scores, the best system among the three systems was System C that scored 477 points, but the scores of all three systems were in the same range.

5. Discussions

By analyzing the results from the Evaluation 1, we have found the following three major problems:

The first problem is that not all examinees did their best. It is often painful to read the machine-translated text, and therefore, the examinees didn't perform well without remuneration. This problem caused the results to be unreliable.

The second is that there were some errors in the manually translated texts, the questions and multiple choices texts. There were some key words translated differently between the manual transla-

tion and the passages translated by the MT systems. For these problems, it was sometimes difficult to select the correct answers, even though the outline of the passage was understood from the machine-translated passages.

The third is that, in contrast to the second problem, some key words in the passages were not translated and appeared in English as in the input text (see *Vibrio parahaemolyticus* in "Appendix B," for example). And the same word in the question, which was manually translated, was left untranslated. The two words, one in the translated passage and another in the question, happened to match, and that helped the examinees to choose the correct answer. More untranslated English words appeared in the texts translated by System B, but System B scored relatively well.

This shows that the untranslated words could sometimes help the readers understand the text. For example, when a word has more than one meaning such as *settlement,* if the system translated *settlement as* "reconciliation" where it actually meant "place newly settled," the context of the passage would be incoherent. If the system couldn't provide the correct translation, it might have been better not to translate the word and leave it in the original language, and expect the readers to know what the word really means. This only applies to certain cases such as English-to-Japanese MT systems for native speakers of Japanese who have basic knowledge of English. If the readers understood the output of an MT system that contains some words not translated, the result was not directly related to the intrinsic performance of the system, because understanding of such translation highly depends on the reader's knowledge of the source language.

In November 1992, the second evaluation, Evaluation 2, was performed on four commercial systems: the same three systems (System A, B, and C) used for Evaluation 1 in June, 1992, and an additional system (System D) that was not available at that time for Evaluation 1. This time, the prizes were implemented. The examinees competed each other to make good scores to receive the prizes. The following tables show the results from Evaluation 2. As shown in the tables, the scores significantly improved. We believe that this improvement was mostly due to implementation of the prizes.

Table 3: Percentages of Correct Answers (Evaluation 2)

	Test 1	Test 2	Test 3	Test 4	Average
System A	72%	67%	48%	85%	68%
System B	75%	82%	63%	83%	76%
System C	73%	80%	58%	92%	76%
System D	73%	92%	42%	88%	74%

Table 4: Estimated TOEFL Scores (Evaluation 1 and 2)

MT Systems	TOEFL Scores (Evaluation 1)	TOEFL Scores (Evaluation 2)	Improved Percent
System A	465	518	+11%
System B	471	550	+17%
System C	477	550	+15%
System D	N/A	544	N/A

6. Concluding Remarks

Throughout our experiments, we found several issues yet to be addressed. In particular, we feel that the following two points, although they may also apply to other psychological experiments, are very important for the future evaluation:

More examinees are required to make the results statistically reliable. In our experiments, for example, in the same unit of 30 questions with the same MT system, one examinee made 28 correct answers (accuracy: 93%), and another examinee made 19 correct answers (accuracy: 63%). The variation of scores by individual examinees indicates that the average scores of 20 examinees per system may not be statistically reliable. We may also need some cut-off procedures to eliminate the worst scores (and perhaps the best scores as well) from consideration.

The prizes and awards evidently affected the examinees' performance. The results from two evaluation (without and with prizes) speak for themselves. The punishment and threat would probably equally effective as well. To read barely translated text is a very painful task. The examinees, therefore, need a strong motive for making the good scores.

Despite those issues, we feel that the evaluation method described in this paper is suitable for comparative evaluation of two or more MT systems for the purpose of skimming the input texts. The objective of our experiments, however, has been to establish an evaluation methodology; not to compare the commercial MT systems.

This evaluation method can also be applied for other language pairs, if a TOEFL-like exam in the source language, and a large number of examinees who are native speakers of the target language are available. In fact, we plan to perform the similar experiments for Japanese-to-English MT systems, by using TOEFL texts professionally translated into Japanese, with the help of a large number of examinees who are native speakers of English.

Our results would be also helpful to other evaluation approaches using multiple choice tests such as the evaluation method initiated by DARPA [4].

References

[1] 'TOEFL Kanzen Enshu," 1987, Yohan Publications, Inc.

[2] M. Nagao, "A framework of a Mechanical Translation between Japanese and English by Analogy Principle," A. Elithorn & R. Banerji (editors), Artificial and Human Intelligence, pages 173 - 180, North Holland, 1984

[3] "Super Course for the TOEFL," 1990, Prentice Hall Press

[4] DARPA, "MT evaluation: basis for future directions," 2-3 Nov. 1992, proceeding of workshop in San Diego

Appendix A

Example of TOEFL Test

The following texts were extracted from a TOEFL guide book [1].

Passage 1

Elizabeth Blackwell was born in England in 1821, and emigrated to New York City when she was ten years old. One day she decided that she wanted to become a doctor. That was nearly impossible for a woman in the middle of the nineteenth century. After writing many letters seeking admission to medical schools, she was finally accepted by a doctor in Philadelphia. So determined was she, that she taught school and gave music lessons to earn money for her tuition.

In 1849, after graduation from medical school, she decided to further her education in Paris. She wanted to be a surgeon, but a serious eye infection forced her to abandon the idea.

Upon returning to the United States, she found it difficult to start her own practice because she was a woman. By 1857 Elizabeth and her sister, also a doctor, along with another female doctor, managed to open a new hospital, the first for women and children. Besides being the first female physician and founding her own hospital, she also established the first medical school for women.

31. Why couldn't Elizabeth Blackwell realize her dream of becoming a surgeon?

    A. She couldn't get admitted to medical school.

    B. She decided to further her education in Paris.

    C. A serious eye infection halted her guest.

    D. It was difficult for her to start a practice in the United States.

32. What main obstacle almost destroyed Elizabeth's chances for becoming a doctor?

A. She was a woman.

B. She wrote too many letters.

C. She couldn't graduate from medical school.

D. She couldn't establish her hospital.

33. How many years elapsed between her graduation from medical school and the opening of her hospital?

A. 8 C. 19

B. 10 D. 36

34. All of the following are "firsts" in the life of Elizabeth Blackwell, except

A. she became the first female physician

B. she was the first woman surgeon

C. she and several other women founded the first hospital for women and children

D. she established the first medical school for women

Passage 2

Vibrio parahaemolyticus is a bacteria that has been isolated from sea water, shell fish, finfish, plankton and salt springs. It has been a major cause of food poisoning in Japan and the Japanese have done several studies on it. They have confirmed the presence of V. parahaemolyticus in the north and central Pacific with the highest abundance in inshore waters, particularly in or near large harbors.

A man named Nishio studied the relationship between the chloride content of sea water and the seasonal distribution of V. parahaemolyticus and concluded that while the isolation of V. parahaemolyticus was independent of the sodium chloride content, the distribution of V. parahaemolyticus in sea water was dependent on the water temperature. In fact it has been isolated in high frequencies during summer, from June to September, but was not isolated with the same frequency in winter.

Within four or five days after eating contaminated foods, a person will begin to experience diarrhea, the most common symptom; this

will very often be accompanied by stomach cramps, nausea, and vomiting. Headache and fever, with *or* without chills, may also be experienced.

55. Which of the following locations would be most likely to have a high concentration of Vibrio parahaemolyticus?

    A. a bay
    B. a sea
    C. the middle of the ocean
    D. sediment

56. The safest time for eating seafood is probably

    A. August
    B. November
    C. July
    D. September

57. The most common symptom of V. parahaemolyticus is

    A. nausea
    B. diarrhea
    C. vomiting
    D. headache and fever

58. The incubation period for this illness is

    A. 2 to 3 days
    B. 3 to 4 hours
    C. 4 to 5 days
    D. several months

59. Nishio's study showed that

    A. the presence of V. parahaemolyticus was dependent on neither the salt content nor the water temperature
    B. the presence of V. parahaemolyticus was dependent on only the salt content
    C. the presence of V. parahaemolyticus was independent of both the water temperature and the salt content
    D. the presence of V. parahaemolyticus was dependent on the water temperature

60. The word cramp in the reading means most nearly

    A. noises
    B. toxicity
    C. severe pain
    D. high temperature

Appendix B

Example of Translation

The following is the translation of the example texts shown in "Appendix A." The passages were translated by the MT systems and the questions and multiple choices texts were translated manually.

Passage 1

エリザベス Blackwell は、1821 年にイギリスで生まれて、彼女が 10 歳だったときにニューヨーク市に海外移住した。
ある日、彼女は医者になりたいと決めた。
それは 19 世紀の中央の女にはほとんど不可能だった。
医学の学校に承認をさがすと多くのレターを書いた後に、彼女は、フィラデルフィアの医者によって最終的には受入れられた。
彼女が学校を教えて、彼女の授業料のためにお金を稼ぐために音楽レッスンを与えて、そう決定することは、彼女だった。
医学の学校からの目盛りの後の 1849 年には、彼女はパリでの彼女の教育を促進すると決めた。
彼女は外科医になりたがっていたが、深刻な目感染は彼女に考えを強制的に捨てさせた。
合衆国に戻るとき、彼女は彼女が女だったので彼女の自己の習慣を始めるのが難しいことがわかった。
1857 年までには、別の雌の医者に伴うまたエリザベスと彼女の姉妹 (医者) は、女達と子供のために新しい病院を 1 番目を置くことがなんとかできた。
彼女の自己の病院の最初の女性内科医と設立であることの他に、また、彼女は女達のために最初の医学の学校を設立した。

31. エリザベス・ブラックウェルは、外科医になるという彼女の夢をなぜ実現することができなかったか
    A. 彼女は医学校に入学を認められなかった
    B. 彼女はパリで勉強する決心をした
    C. 彼女は、重い目の病気のため、それを求めることができなくなった
    D. 彼女がアメリカで開業するのは困難だった

32. エリザベスは、どんな障害のために医者になる機会を失いかけたか
    A. 彼女は女性だった
    B. 彼女は手紙を多く書きすぎた
    C. 彼女は医学校を卒業できなかった
    D. 彼女は自分の病院を設立できなかった

33. 医学校を卒業し、開業するまでの間に何年が経過したか

A. 8　　C. 19

B. 10　　D. 36

34. 次のうち、エリザベス・ブラックウェルの人生において「最初」でなかったのはどれか

A. 彼女は最初の女医になった

B. 彼女は最初の女外科医になった

C. 彼女と他の数人の女性は、女性と子供達のための最初の病院を設立した

D. 彼女は女性のための最初の医学校を設立した

Passage 2

Vibrio parahaemolyticus は、海水、殻魚、finfish、プランクトンおよび塩春から孤立しているバクテリアである。
日本と日本人の食中毒の主要な原因したそれのいくつかの調査それだった。
それらは、Vの出席を確認した。
沿岸の水で一番高い多量のある北と主要な太平洋、の parahaemolyticus は、特に大きいの中または近くでかくまう。
Nishio と名づけられる男は、海水の塩化物内容とVの季節特有の分配の間で関係を勉強した。
parahaemolyticus そしてVの孤立の間それを決定されると。
parahaemolyticus は、塩化ナトリウム内容Vの分配に、依存しなかった。海水の parahaemolyticus は、水温度を頼った。
実際、それは、6月からの9月まで夏の間に高い頻度で孤立している、しかし、冬の同じ頻発で孤立していなかった。
4ないし5日以内に、汚染された食品を食べた後で、人は、下痢一番共通の徴候を、経験し始めるだろう；これが胃けいれん、吐き気および嘔吐することをほんとうにしばしば伴っているだろう。
頭痛と熱も、または悪寒なしに、で経験されるかもしれない。

55. 次の場所のうち、Vibrio parahaemolyticus がもっとも集中する傾向にあるのはどこか

A. 湾　　C. 遠洋

B. 海　　D. 堆積物

56. 海産物を食べるのに一番安全な時期は次のうちいつか

A. 8月　　C. 7月

B. 11月　　D. 9月

57. V. parahaemolyticus でもっとも現われやす症状は次のうちどれか

A. 吐き気　　C. 嘔吐

B. 下痢　　D. 頭痛と熱

58. この病気の潜伏期間はどれくらいか

A. 2 ～ 3 日　　C. 4 ～ 5 日

B. 3 ～ 4 時間　　D. 数カ月

59. 西尾の研究は、次のどれを示していたか

A. V. parahaemolyticus の存在は塩分濃度にも水温にも依存していなかった

B. V. parahaemolyticus の存在は塩分濃度だけに依存していた

C. V. parahaemolyticus の存在は水温と塩分濃度のどちらにも依存していなかった

D. V. parahaemolyticus の存在は水温に依存していた

60. この文章の中の「けいれん」の意味に最も近いものは

A. 雑音　　C. ひどい苦痛

B. 毒性　　D. 高温