Evaluating Machine Translation Systems with Second Language Proficiency Tests

Takuya Matsuzaki[†‡] **Akira Fujita**[‡] **Naoya Todo**[‡] **Noriko H. Arai**[‡]
[†]Dept. of Electrical Engineering and Computer Science, Nagoya University
matuzaki@nuee.nagoya-u.ac.jp
[‡]Information and Society Research Division, National Institute of Informatics
{a-fujita, ntodo, arai}@nii.ac.jp

Abstract

A lightweight, human-in-the-loop evaluation scheme for machine translation (MT) systems is proposed. It extrinsically evaluates MT systems using human subjects' scores on second language ability test problems that are machine-translated to the subjects' native language. A large-scale experiment involving 320 subjects revealed that the context-unawareness of the current MT systems severely damages human performance when solving the test problems, while one of the evaluated MT systems performed as good as a human translation produced in a context-unaware condition. An analysis of the experimental results showed that the extrinsic evaluation captured a different dimension of translation quality than that captured by manual and automatic intrinsic evaluation.

1 Introduction

Automatic evaluation metrics, such as the BLEU score (Papineni et al., 2002), were crucial ingredients for the advances of machine translation technology in the last decade. Meanwhile, the shortcomings of BLEU and similar n-gram proximity-based metrics have been pointed out by many authors including Callison-Burch et al. (2006). The main criticisms include: 1) unreliability in evaluating short translations, 2) non-interpretability of the scores beyond numerical comparison, and 3) bias towards statistical MT systems.

Manual evaluation of translation quality is more reliable in many regards, but it is costly. Furthermore, it is not necessarily easy to *analyze* the characteristics of MT systems based solely on the evaluation results such as a 5-point scale evaluation of adequacy/fluency and a ranking of the outputs of different systems.

A remedy for some of the above-raised issues is task-based evaluation of MT systems (Jones et al., 2005; Voss and Tate, 2006; Laoudi et al., 2006; Jones et al., 2007; Schneider et al., 2010; Berka et al., 2011), which measures the human performance in a task such as information extraction from a machine-translated text. The main burden of conducting task-based evaluation is also its cost; the development of a sizable amount of test materials and the gathering of appropriate human subjects is time consuming and expensive.

This paper proposes to utilize second-language proficiency tests (SLPTs), such as TOEIC, as the source of the specimens for extrinsic evaluation of MT systems. For evaluating, e.g., English-to-Japanese MT systems, a set of English test problems is translated by the systems and the translation qualities are evaluated by the test scores achieved by native Japanese speakers on the translated problems.

In many languages, a large collection of SLPT problems is available. More than 130 standardized tests for 32 languages are listed in the English Wikipedia page for 'List of language proficiency tests' as of April 30, 2015. They are carefully designed to evaluate various aspects of language ability with objective criteria. We can thus obtain an easy-to-use test set that focuses on a certain aspect of MT system performance by appropriately choosing the problem types and levels. Moreover, SLPTs are primarily designed to assess the test-takers' language ability but not their general intelligence. Hence, as evidenced later in the paper, the proposed scheme is expected to be robust against the heterogeneity of the subjects, as long as they are native speakers of the target language. This is a desirable property for conducting a large-scale experiment, possibly with crowdsourcing.

In the current paper, we utilize a typical format of multiple-choice dialogue completion problems (Figure 1). The subjects are given a machine-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 145–149,
Beijing, China, July 26-31, 2015.

INSTRUCTION
Choose the most suitable utterance for the blank in the following dialogue from choices 1, 2, 3, and 4.
DIALOGUE
A: Hello. Can I help you?
B: Yes. [**BLANK**]
A: I'm sorry, I can't find that name on the reservation list.
B: Oh, really? Then give me a new reservation, please.
OPTIONS
1. I'd like to make a reservation for Flight 502.
2. I have a reservation under the name Hashimoto.
3. I'm sure you can find my name on the list.
4. I wonder if you could tell me how to make a reservation.

Figure 1: Example of multiple-choice dialogue completion problem

translated conversation and asked to choose an appropriate utterance from several options, which are also machine-translated, to fill in a blank in the conversation.

We evaluated four translation methods in the experiment including both machine-translation and manual-translation. The extrinsic evaluation revealed that one of the MT systems is comparable to the human translation produced by randomly presenting the individual sentences to the translator without any context, but the translation produced by the best MT system is still far worse than that produced by a human translator working on the entire dialogue at once. Furthermore, we examined the relations between the extrinsic metric based on the subjects' scores and various intrinsic metrics including automatic scores such as the BLEU score and manual evaluation. The test material is available on request for research purposes.

2 Method

2.1 Overview of Experiment

We extrinsically evaluated four different translations of the same material, namely multiple-choice dialogue completion problems taken from second language ability tests. The original problems were in English, and we translated them into Japanese. Two of the translations were produced by MT systems, and the other two were produced by a human translator with and without considering the contexts of the individual sentences in the dialogues. The human subjects solved the translated problems without knowing whether a machine or a human produced them. Finally, the translation quality was evaluated based on the rate of correct answers given by the human subjects.

2.2 Participants

The subjects of the experiment included 320 Japanese junior high school students (12-15 years old) from two schools (schools A and B). The participants from school A consisted of 80 first-year students, 80 second-year students, and 78 third-year students. All the students from school B (82 students) were first-year students. Thus, the participants had varying levels of English and scholastic abilities. We will examine the effect of these factors on the experimental results later in the paper.

2.3 Materials

All the problems used in the experiment consisted of a short conversation between two people, where part of an utterance is hidden. The subject was presented with four options and asked to complete the dialogue with the most appropriate one.

We randomly extracted 40 English dialogue completion problems from mock National Center Test for University Admissions conducted by one of the largest preparatory schools in Japan. In the extracted problems, the number of utterances in one dialogue ranged from two to four, with each utterance consisting of one to three sentences, and an option including one or two sentences. All 40 problems contained 327 sentences.

2.4 Translation Systems

The English dialogue completion problems were translated by four methods: [1]

G**:** Automatic translation by Google Translate[2]

Y**:** Automatic translation by Yahoo Translate[3]

H_S**:** Human translation produced by providing individual sentences from the problems to a translator in random order

H_O**:** Human translation produced by a translator working on the entire dialogue at once

The subscripts of H_S and H_O stand for the translations of the sentences in "shuffled order" and "original order", respectively. The translations by H_S were created by first preparing a file containing all the sentences from the 40 problems in a randomized order and then asking a translator to translate the file sentence-by-sentence, without assuming any specific context. H_S thus provides

[1] The two MT results were produced on June 11th, 2014.
[2] https://translate.google.co.jp/?hl=ja
[3] http://honyaku.yahoo.co.jp/

an estimate of the performance upper-bound of the current MT systems since most current systems translate each sentence individually.

We asked three native Japanese speakers who are fluent in English to first produce the sentence-by-sentence translations by method H_S and then translate all the dialogue problems in the normal way (i.e., by H_O). We randomly chose one of the translators and used his translations as the test material that the subjects solved. The other human translations were used as the reference translations for the automatic evaluation.

2.5 Procedure

Each subject was given 12 different problems that consisted of an equal number (3) of translated problems produced by the four translation methods. Although the sets of problems were different among the subjects, they were designed such that the number of subjects who solve each translated problem was roughly the same. Each subject was given 12 sheets of paper, each of which showed a problem and its answer choices, and was given one minute to complete each problem.

2.6 Extrinsic Evaluation Metric

The translation systems were evaluated by the average of the rate of correct answers made on the translated problems. Let $P = \{p_i\}$ be the set of original problems and $M(p)$ be the translation of problem p produced by method M. The correct answer rate (CAR) on $M(p)$ is defined as:

$$\text{CAR}_M(p) = \frac{\text{\# of subjects that correctly answered } M(p)}{\text{\# of subjects who solved } M(p)}.$$

The extrinsic evaluation score of translation method M is the average of CAR over P:

$$\text{Avg-CAR}_M = \frac{1}{|P|} \sum_{p \in P} \text{CAR}_M(p).$$

2.7 Intrinsic Evaluations

Automatic Evaluation Metrics We also evaluated the translation quality using BLEU, BLEU+1 (Lin and Och, 2004), RIBES (Isozaki et al., 2010), and TER (Snover et al., 2006). We prepared two sorts of reference translations: Ref_S and Ref_O. Ref_S consisted of two manual translations of the 40 problems produced by method H_S. Ref_O consisted of three manual translations produced in the normal way, i.e., by H_O.

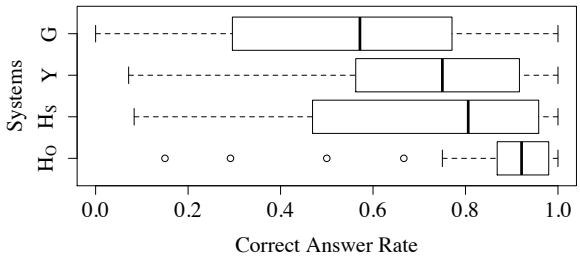

Figure 2: Boxplots of Correct Answer Rates for 40 Problems

Human Evaluation Five native Japanese speakers ranked the translations by the four systems for each of the 40 problems. They were shown the translations of a problem by the four methods with its source problem in English and asked to give a relative ranking among them, such as "$G < Y < H_S = H_O$." This method was adapted from the manual evaluation conducted in the recent WMT workshops (Callison-Burch et al., 2010). The relative ranking was broken down into six ($= {}_4\text{C}_2$) binary relations. For each relation "$A > B$" found in the broken-down relations, one point was added to system A. The final ranking among the systems for a problem was determined by the total points.

3 Results and Discussion

3.1 Preliminary Analysis: Robustness against the Heterogeneity of the Human Subjects

We divided the participants from school A into three groups according to grade level, and then examined the differences in the rate of correct answers for each problem among each group. We also compared the correct answer rates between the participants in the 1st grade at schools A and B. The two-way analysis of variance (ANOVA) revealed that the grades and schools had no significant effect on the correct answer rate for 38 out of the 40 problems ($p > 0.05$). The results showed that the participants' grade levels and scholastic abilities (including English ability) did not affect the test results.

3.2 System-level Evaluation

We first present the system-level evaluation results for the four translation methods. Figure 2 shows the min/max and the quartiles of the correct answer rates (CARs) for the 40 problems translated by each system. The averages of the correct answer rates are 0.524, 0.696, 0.693, and 0.875 for each translation system G, Y, H_S, and H_O, re-

Reference	Metrics	G	Y	H_S	H_O
Ref_O	BLEU	22.04	20.33	40.30	47.43
	BLEU+1	22.08	20.37	40.33	47.46
	RIBES	67.80	69.43	78.16	82.42
	TER	41.72	43.66	27.47	24.14
Ref_S	BLEU	27.53	23.63	41.24	30.69
	BLEU+1	27.56	23.67	41.27	30.73
	RIBES	73.61	73.63	80.18	70.59
	TER	36.51	39.52	27.60	31.51
Avg-CAR		0.524	0.696	0.693	0.875

Table 1: Automatic Evaluation Scores and Average Correct Answer Rate

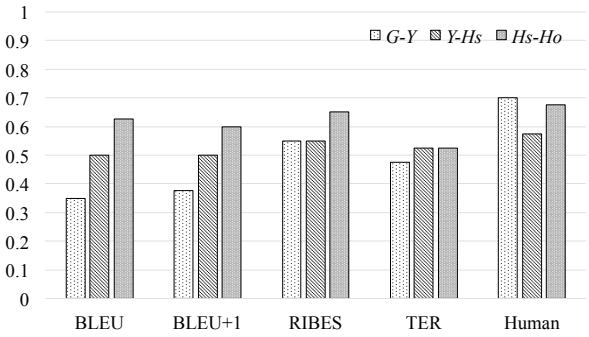

Figure 3: Agreement Rates between Intrinsic Evaluation Metrics and Correct Answer Rate

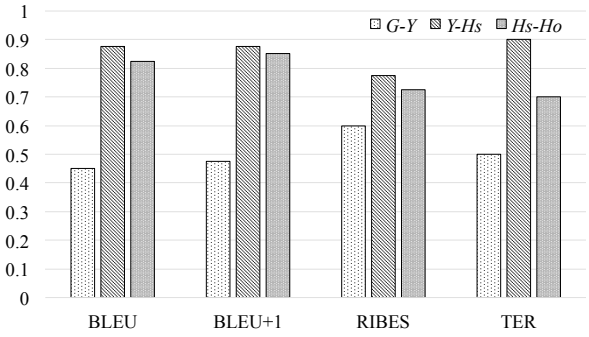

Figure 4: Agreement Rates between Automatic Evaluation Metrics and Human Evaluation

spectively. We conducted a pairwise t-test on each adjacent set (G-Y, Y-H_S, and H_S-H_O) for the CARs and found a statistically significant difference ($p < 0.05$) between G and Y and H_S and H_O but not between Y and H_S ($p = 0.954$).

Table 1 lists the five automatic evaluation scores for each translation method measured against the two reference translation sets. The averages of the CARs over the 40 problems are also listed in the bottom row of the table. There are several noticeable facts. First, despite the significantly better average CAR for Y over G, BLEU, BLEU+1, and TER prefer G to Y. Second, while the average CARs for Y and H_S are almost equal, there are large differences between their automatic evaluation scores across all metrics. Third, a comparison of the corresponding automatic evaluation scores using Ref_S and Ref_O reveals that G, Y, and H_S are more similar to the manual translations that were produced without referring to the contexts of the individual sentences than those produced taking the contexts into consideration. This is not surprising. However, the large difference in the correct answer rates for H_S and H_O suggests that ignorance of the context in the current MT systems severely degrades the comprehensibility of the translations of texts like daily conversations.

3.3 Agreement between Intrinsic and Extrinsic Evaluation Metrics

We examined how often an intrinsic metric correctly predicts the difference of the subjects' test performance on a problem. Specifically, for two translation methods A and B, we say an intrinsic metric M agrees with the CAR by the subjects on problem p_i iff metric M scores A's translation of p_i (= $A(p_i)$) better than B's translation (= $B(p_i)$) and the CAR is higher on $A(p_i)$ than on $B(p_i)$. The rate of agreements is the fraction of the problems on which M and CAR agree. The agreement between two intrinsic metrics is defined similarly.

Figure 3 shows the rates of agreements between the automatic metrics and CARs and between the human evaluation and CARs. As Figure 3 shows, all the agreement rates between the automatic metrics and CARs were less than 0.65. When considering a random baseline of 0.5, we may conclude that the automatic metrics are not very good predictors of the CARs. This is unfortunate since the CARs directly indicate the comprehensibility of the translated dialogues. The disagreements cannot be attributed only to the unreliability of automatic metrics on short translations. Figure 4 shows the rate of agreements between the automatic metrics and the human evaluation. As Figure 4 shows, BLEU, BLEU+1, and TER agree with human evaluation on nearly 90% of the problems when comparing Y and H_S.

The human evaluation agrees with the CAR slightly better than the automatic metrics. However, the agreement rates are still less than 0.7 for all pairs of compared systems. These findings suggest that there is an inherent discrepancy between the assessment of the overall translation quality of

a problem and the CAR. It is presumably because the CAR can be critically damaged by a subtle translation mistake that spoils a coherent understanding of a dialogue.

4 Conclusion and Future Work

We have presented the results of an experiment, in which machine- and human-translated second language proficiency test (SLPT) problems were used for extrinsic evaluation of the translation quality. Comparison of four translation methods revealed, most notably, the crucial importance of considering contexts of individual sentences in translating dialogues. The analysis on the experimental results suggests that the extrinsic evaluation based on SLPT problems captures a different dimension of translation quality than the manual/automatic intrinsic metrics. The robustness against the heterogeneity of human subjects and the abundance of existing SLPT problems enable easy adaption of the proposed evaluation scheme in addition to the traditional intrinsic evaluations. Our future work includes experiments with other types of SLPT problems that focus on different aspects of translation quality and language understanding.

Acknowledgments

The authors are grateful to all the participants in the experiments for their time and patience and also grateful to Yoyogi Seminar for their allowance to use their problems in the experiments. This study is conducted as a part of the Todai Robot Project (`http://21robot.org/?lang=english`).

References

Jan Berka, Martin Černỳ, and Ondřej Bojar. 2011. Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77–86.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 249–256.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)*, pages 944–952.

Douglas Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. 2005. Measuring translation quality by testing english speakers with a new defense language proficiency test for arabic. In *Proceedings of the 2005 International Conference on Intelligence Analysis*.

Douglas Jones, Martha Herzog, Hussny Ibrahim, Arvind Jairam, Wade Shen, Edward Gibson, and Michael Emonts. 2007. Ilr-based mt comprehension test with multi-level questions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007); Companion Volume, Short Papers*, pages 77–80.

Jamal Laoudi, Calandra R. Tate, and Clare R. Voss. 2006. Task-based mt evaluation: From who/when/where extraction to event understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, pages 2048–2053.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 501–507.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 311–318.

Anne H. Schneider, Ielka van der Sluis, and Saturnino Luz. 2010. Comparing intrinsic and extrinsic evaluation of mt output in a dialogue system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-2010)*, pages 329–336.

Matthew Snover, Bonnie Dorr, Richard Schwaltz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA-2006)*, pages 223–231.

Clare R. Voss and Calandra R. Tate. 2006. Task-based evaluation of machine translation (mt) engines: Measuring how well people extract who, when, where-type elements in mt output. In *In Proceedings of 11th Annual Conference of the European Association for Machine Translation (EAMT-2006)*, pages 203–212.