Translation Errors and Incomprehensibility: A Case Study using Machine-Translated Second Language Proficiency Tests

Takuya Matsuzaki[†‡], **Akira Fujita**[‡], **Naoya Todo**[‡], **Noriko H. Arai**[‡]
[†]Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603, JAPAN
[‡]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8634, JAPAN
matuzaki@nuee.nagoya-u.ac.jp, {a-fujita, ntodo, arai}@nii.ac.jp

Abstract

This paper reports on an experiment where 795 human participants answered to the questions taken from second language proficiency tests that were translated to their native language. The output of three machine translation systems and two different human translations were used as the test material. We classified the translation errors in the questions according to an error taxonomy and analyzed the participants' response on the basis of the type and frequency of the translation errors. Through the analysis, we identified several types of errors that deteriorated most the accuracy of the participants' answers, their confidence on the answers, and their overall evaluation of the translation quality.

Keywords: Machine Translation, Evaluation, Error Analysis

1. Introduction

What level of "quality" is required for machine translation (MT) systems should change depending on who uses them in what situation. If it is used as a web page translator, even fragmentary information will be helpful. However, if it is used as an automatic interpreter on smartphones, it is a totally different type of matter. It is expected to convey correct information between two people who do not understand each other's language, sometimes in critical situations. In other words the measurement of the quality of MT systems should include not only the intrinsic metrics such as accuracy and fluency but also how often a user can complete his/her purpose in using it.

In our previous paper, we proposed a light weight, human-in-the-loop extrinsic evaluation scheme (Matsuzaki et al., 2015), where MT systems are evaluated by human subjects' scores on the second language ability tests translated to the subjects' native language by MT. Specifically, we used dialogue completion questions (Figure 1) as the test material. Dialogues often involve linguistic phenomena that are not frequent in written text, such as interrogatory sentences, imperative sentences, and ellipsis. It turned out that our evaluation captured a different dimension of translation quality than that captured by manual and automatic intrinsic evaluation.

In this paper, we report on another experiment involving 795 human participants, and scrutinize what kinds of errors hinder the users' comprehension of the dialogues more often than others. Although MT-mediated conversation is already in the scope of development, automatic metrics such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) do not tell us which errors are more harmful than others. The translation errors in the test materials were manually classified according to an error taxonomy. Then, the subjects' responses on the translated questions were analyzed on the basis of the type and frequency of the translation errors. Through the analysis, we identified several types of errors that deteriorated most the subjects' performance, their confidence on the answers, and the overall evaluation of the translation quality.

INSTRUCTION
Choose the most suitable utterance for the blank in the following dialogue from choices 1, 2, 3, and 4.
DIALOGUE
A: Jack, I just finished washing your school uniform, and found your cellphone in the washing machine. It's broken!
B: Oh, no. I have to call Bob now.
A: That's not the point! I just bought it for you last week!
B: Oh, yeah. I'm so sorry. But Mom, how am I going to call him?
A: [**BLANK**] We'll talk about your carelessness later.
OPTIONS
1. Buy him a new phone.
2. I'll call you soon.
3. Just use my phone.
4. Tell him to wait for me.

Figure 1: Multiple-choice dialogue completion question

2. Experimental Material and Procedure

2.1. Material

All the test material used in this study were dialogue completion questions taken from the English tests in National Center Test for University Admission (NCTUA)[1]. A question in this format consists of a short conversation between two people, where one of the utterances is hidden ([BLANK] in the figure). One has to choose an appropriate utterance which fills the blank out of four choices.

We collected 200 past NCTUA dialogue completion questions and translated them to Japanese using three MT systems: Google translation[2], Yahoo! translate[3], and a system developed by National Institute of Information and Communications Technology (NICT) in Japan[4][5]. The translated questions were firstly classified into three groups:

[1]NCTUA is a national standardized test in Japan.
[2]`https://translate.google.co.jp/?hl=ja`
[3]`http://honyaku.yahoo.co.jp`
[4]`https://mt-auto-minhon-mlt.ucri.jgn-x.jp`
[5]All the MT results were produced on May 30th, 2015.

- questions including serious translation errors
- questions including minor translation errors
- questions translated mostly correctly

The questions are then sorted in the descending order of the number of the systems that made serious translation errors on the questions. Ties were broken by the number of systems that made minor errors. We then chose 50 questions ranked highest in the list as the test material while skipping such questions that involve linguistic issues in English but not in Japanese (e.g., subject-verb agreement). They included 463 sentences and 2,557 words in total.

The test material is thus not a representative or balanced example of English conversation (if such a thing exists) but designed to investigate how different types of translation errors affect the human subjects' task performance. Please refer to our previous work for the system performance comparison based on randomly chosen question set (Matsuzaki et al., 2015).

In addition to the three MT results, we prepared two kinds of human translations. One was produced by translating all the sentences in the test material in a randomized order (hereafter Human-Shuffle). The other was produced by translating the entire dialogue at once (Human-Original). The Human-Shuffle translation was used to investigate the effect of translation errors purely caused by the ignorance of the context (as is done in most current MT systems).

2.2. Procedure

795 high school students participated in the experiment as the subjects. We prepared 25 different combinations of translated questions, each of which includes ten questions. The question sets were designed so that

- a student answers ten different questions,
- a question set includes no two translations of the same question and includes the same number (two) of translations produced by each of the five systems, and
- approximately the same number of answers are obtained for each translated question.

On average, 31.8 answers were collected for each pair of question and translation system.

The subjects answered to each question in one minute. They were also asked to indicate their confidence on their answer on each question in three levels:

A: I am fully confident about my choice.
B: I am not fully confident about my choice.
C: I do not have any confidence at all.

After answering each question, the subjects were asked to evaluate the MT system that produced the translation:

A: This MT system will be useful.
B: This MT system may be useful in some situation.
C: This MT system would not be useful in most situation.

We converted A, B, and C to 3, 2, and 1 respectively for the purpose of quantitative analysis described in §4.

- Missing words
- Word order
- Incorrect words
  - Wrong lexical choice
  - Disambiguation
  - Extra words
  - Idioms
  - Incorrect forms
    - Inflection and tense of predicates
    - Animacy (human or animal / object)
  - Not translated (just transliterated)
- Incorrect dependency
  - Between a predicate and an NP argument
    - With Nominative case
    - With Accusative case
    - With Genitive case
    - Other cases
  - Between two clauses
    - Incorrect translation of conjunctions
    - Incorrect translation of other connective expressions
- Grammatical property
  - Incorrect Voice (active or passive)
  - Incorrect Mood (declarative, interrogative, or indicative)
  - Incorrect Polarity (affirmative or negative)
  - Incorrect Modality (degree of probability, obligation, necessity, evidentiality, etc.)
  - Violation of very basic word order
- Semantic
  - Sentiment polarity
  - Ellipsis

Figure 2: Error classification taxonomy

3. Translation Error Classification

3.1. Error Taxonomy

The translation errors in the test material were classified according to the taxonomy shown in Figure 2. First three categories ("missing words", "word order", and "incorrect words") are basically the same as those in Vilar et al. (2006)'s error classification; We added a sub-category to "incorrect form" that is applied to a wrong choice of the animacy (human or animal / inanimate thing) of pronoun and its agreement with a Japanese copula verb (*iru* / *aru*).

"Incorrect dependency" stands for a wrongly expressed grammatical relation between a predicate and an NP argument or between two clauses. They are further subcategorized according to the grammatical cases of the NP arguments and the type of grammatical devices that connect two clauses (conjunctions or postpositions and affixes).

"Grammatical property" errors are applied to the mistranslations that result in wrong alternation of some grammatical property of a sentence (voice / mood / polarity / modality) or a violation of very basic word order (e.g., a proper Japanese sentence never begins with a postposition).

"Semantic" errors include two frequent error types found in the test material. "Sentiment polarity" errors are applied to the cases where the translated utterance expresses different sentiment polarity (being positive / negative about an event or an object) than the original utterance. For instance, in-

terjections such as "oh!" can express both positive and negative sentiment. When they are translated to Japanese expressions with the opposite polarity, the dialogue becomes very unnatural and may cause misunderstanding.

An "Ellipsis" error is caused by a wrong supplementation for an omitted phrase. For instance, VP-ellipsis in an answer to a Yes-No question is often supplemented wrongly with a random phrase, as in the following example (Question ID 184, translation by NICT):

EN:	"You didn't tell her?"		– "No, I didn't."
JP:	Kanojo-ni	iwa-naka-tta-nda?	– Tabe-masen.
	she-DAT	*say*-NEG-PAST-Q	*eat*-NEG

("You didn't tell her, did you?" – "No, I don't eat it")

3.2. Error Classification Process

Two annotators parallelly classified the errors in all the translated questions. The annotators were provided with the translated questions in which the blanks in the dialogue were filled with the translations of the correct choices. This is because putting the wrong choices in the blank makes the whole dialogue incomprehensible or unnatural, and thus makes the error classification more difficult.

We asked the annotators to firstly translate the original English questions so that the translation is as close to the systems' output as possible. The annotators then compared their translations and the systems' output and classified the differences according to the taxonomy. This two-step process was to make the classification less subjective. An error could be classified to more than one leaf categories (but not into the same top/mid-level categories).

The classification was based only on the translated questions; The annotators carried out the process without knowing which MT system produced the translation or how many subjects answered the question correctly. The annotators' translations were also used as the reference translations for the calculation of the automatic metrics (§4.3).

4. Experimental Results

4.1. Error Profile

Table 1 shows the profile of the errors in the test material. The numbers are the percentages of the sentences that include at least one error that was categorized to each of the error types (averaged over the results by the two annotators). Although the test materials are not random samples, the error profiles still seem to reflect the basic architecture of the three MT systems. Comparing to the two statistical systems (i.e., Google and NICT), the rule-based one (i.e., Yahoo!) makes far less errors categorized in "Predicate-argument deps." and "Grammatical property," which are mainly related to grammar and syntax. Meanwhile, the frequencies of the errors related to semantics and discourse are comparable across the three MT systems ("Incorrect words - Disambiguation" and "Semantic").

	Google	NICT	Yahoo!	Shuffle
Missing words	1.9	15.2	1.0	3.5
Word order	5.0	1.4	1.7	0.2
Incorrect words	56.9	35.6	36.9	4.3
Wrong lexical choice	0.7	2.2	0.2	0.0
Disambiguation	30.1	16.1	22.3	2.8
Extra words	8.0	6.2	2.8	0.0
Idioms	19.0	8.5	9.9	0.7
Inflection and tense	10.9	9.9	7.3	0.9
Animacy	2.2	2.1	1.9	0.0
Transliterated	5.5	1.6	1.9	0.0
Predicate-argument deps.	15.6	19.0	4.0	1.4
Nominative case	3.8	10.9	0.9	0.5
Accusative case	6.7	5.4	0.9	0.5
Genitive case	1.2	1.0	0.0	0.0
Other cases	5.4	3.6	2.2	0.3
Deps. between clauses	10.2	6.4	4.0	0.5
Conjunctions	4.0	3.1	3.5	0.5
Other connectives	6.4	3.3	0.5	0.0
Grammatical property	13.5	14.2	4.3	1.2
Voice	1.4	0.5	0.3	0.0
Mood	6.6	3.1	1.0	0.5
Polarity	1.2	1.2	0.3	0.2
Modality	3.8	3.8	2.6	0.3
Basic word order	1.4	6.2	0.0	0.2
Semantic	5.9	6.4	5.9	3.1
Sentiment polarity	2.9	3.1	3.3	1.7
Ellipsis	3.3	3.3	2.6	1.4

Table 1: Profile of the translation errors in the test material (percentages of the sentences that include at least one error classified to the error categories)

	Independent variables		
Coefficients	y_{RCA}	y_{conf}	y_{eval}
(Intercept)	-0.10***	-0.16***	-0.27***
b_{Google}	0.03	-0.13	-0.31***
b_{NICT}	-0.01	-0.21**	-0.35***
$b_{\text{Yahoo!}}$	0.01	-0.16**	-0.29***
Missing Words	-2.07**	-1.87*	-0.68
Word Order	-0.60	-2.53	-1.03
Incorrect words	-2.16***	-2.99***	-2.23***
Pred-arg deps.	0.33	-1.29	-1.10
Deps. btwn clauses	-1.55	-3.81**	-1.58
Grammatical property	-1.52*	-2.95**	-2.59***
Semantic	-2.74**	-3.12**	-3.18***
Adjusted R^2	0.366	0.554	0.613

(Legend: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$)

Table 2: Estimated coefficients of the regression models with the number of errors classified to the top categories as the independent variables

4.2. Effect of the Errors on Extrinsic Metrics

We conducted regression analyses to examine the effect of different error types on the subjects' responses. The dependent variable y_{RCA} stands for the rate of correct answers in our linear regression model. We conducted two more regression analyses where we set the subjects' confidence on their answers (y_{conf}) and their evaluation of the MT systems (y_{eval}) for the dependent variables. The parameters were estimated using the fifty questions translated by Google, NICT, Yahoo!, and Human-Shuffle all at once, i.e., the number of samples was 200.

The dependent values were shifted by the corresponding

Coefficients	Independent variables y_{RCA}	y_{conf}	y_{eval}
(Intercept)	-0.09***	-0.15***	-0.28***
b_{Google}	0.03	-0.13	-0.32***
b_{NICT}	-0.04	-0.23**	-0.34***
$b_{\text{Yahoo!}}$	0.03	-0.14*	-0.30***
Missing words	-2.50***	-1.65	-0.27
Word order	-0.05	-1.87	-0.90
Incorrect words			
Wrong lexical choice	-3.63	-8.77**	-4.84
Disambiguation	-2.89***	-3.82***	-2.27***
Extra words	-1.82	0.82	-0.77
Idioms	-1.66*	-4.42***	-3.45***
Inflection and tense	-1.46	-3.59**	-2.97**
Animacy	-2.10	0.49	-0.92
Not translated	-1.08	0.93	1.11
Predicate-argument deps.			
w/ Nominative case	1.00	-0.88	-1.43
w/ Accusative case	1.04	-0.91	-0.55
w/ Genitive case	-2.11	-5.37	-6.93*
w/ other cases	0.37	-0.76	1.02
Deps. between clauses			
via Conjunctions	-2.34	-4.64**	-0.84
via Other connectives	-0.79	-1.88	-1.18
Grammatical property			
Incorrect Voice	4.93	-2.46	-2.98
Incorrect Mood	-3.21**	-4.83**	-4.31**
Incorrect Polarity	0.34	-5.72	-1.62
Incorrect Modality	-5.26***	-2.71	-1.03
Very basic word order	1.57	-1.80	-3.67*
Semantic			
Sentiment polarity	-3.47*	-4.15*	-2.88
Ellipsis	-1.73	-1.54	-2.62*
Adjusted R^2	0.37	0.56	0.61

(Legend: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$)

Table 3: Estimated coefficients of the regression models with the number of errors in the leaf categories as the independent variables

values on the human translation (Human-Original) to reduce the effect of the inherent difficulty of the questions. For instance, y_{RCA} for question q translated by Google was set to:

$$y_{\text{RCA}} = \text{RCA}(\text{Google}, q) - \text{RCA}(\text{Human-Original}, q)$$

where RCA(s, q) stands for the rate of correct answers on question q translated by system s.

The independent variables are the numbers of translation errors categorized according to the taxonomy. We normalized them by the total number of words in a question to reduce the effect of the length of the questions. In addition, we used three binary (i.e., 1 or 0) dependent variables b_{Google}, b_{NICT}, and $b_{\text{Yahoo!}}$, that signify the MT systems that produced the translation. The coefficients of the binary variables represent different intercepts (i.e., the predicted dependent value on a question including no error) for the three MT systems, while the constant term of the model represents the intercept for Human-Shuffle.

Table 2 lists the estimated coefficients of the models where independent variables are the numbers of the errors aggregated at top-most error categories. The adjusted R^2 values

System	vs RCA	vs CONF	vs EVAL
Google	0.30	0.33	0.36
NICT	0.37	0.66	0.56
Yahoo!	0.54	0.59	0.64
Human-Shuffle	0.44	0.53	0.50
All	0.54	0.70	0.71

Table 4: Correlation coefficients between BLEU and subjects' responses

show that the models for the confidence on the answers and the system evaluation fit the data modestly, while the rate of correct answers cannot be fully explained by the current regression model. Nonetheless, we can see some types of errors do have more effect on the rate of correct answers: the coefficients for y_{RCA} suggest that missing word errors, incorrect word errors, and semantic errors are more harmful than other types. Coefficients for y_{conf} and y_{eval} suggest that they are affected by a wider variety of error types, especially by errors related to grammar and syntax.

We also conducted the regression analysis where we set the number of errors in the leaf-level categories for the independent variables. Table 3 shows the estimated coefficients. The results for y_{RCA} reveals that incorrect mood, incorrect modality, and sentiment polarity errors have high negative impact on the subjects' test performance though they are rare in our sample set. The error profile in Table 1 also suggests, overall, word disambiguation errors and literal translations of idioms damaged the subjects' performance more often and severely. The coefficients for y_{conf} and y_{eval} show they are affected by slightly different types of errors than y_{RCA}, such as the large negative effect of wrong lexical choice on y_{conf}.

4.3. Automatic Intrinsic Evaluation and Extrinsic Metrics

How well does an automatic intrinsic evaluation metrics such as BLEU (Papineni et al., 2002) predict the subjects' responses? Figure 3 presents the scatterplots of BLEU score of translated questions and the rate of correct answers (RCA), the average of the subjects' confidence (CONF), and system evaluation (EVAL). As the figures and the correlation coefficients (r) show, the BLEU score correlates well with the subjects' confidence and the system evaluations, but only modestly with RCA.

Table 4 summarizes the correlation coefficients calculated only on the questions translated by a single system (i.e., #samples = 50) and on all of them (i.e., #samples = 200). The table reveals that the BLEU score does not correlate well with all of RCA, CONF, and EVAL when calculated only on the translations by Google. It also shows the BLEU score correlates modestly with RCA only on the translations by Yahoo! and Human-Shuffle.

4.4. Qualitative Analysis

Finally, we examined the translated questions on which more than 90% of the subjects chose wrong answers (ten translated questions in total). We identified five types of errors that are supposed to be the main reasons of the wrong

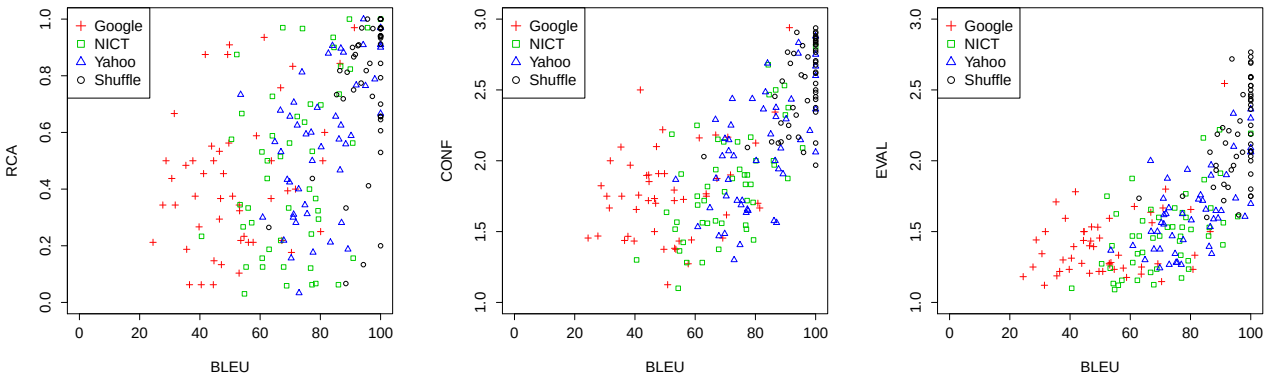

Figure 3: Scatterplot of BLEU score and evaluation metrics. Left: BLEU vs. RCA (r = 0.54); Middle: BLEU vs. CONF (r = 0.70); Right: BLEU vs. EVAL (r = 0.71).

answers:

1. Incorrect translation of imperative sentences
2. Incorrect translation of modality
3. Wrongly supplied Japanese phrases for omitted phrases in the English sentences
4. Literal translations of the sentences including omitted phrases with no supplementation
5. Wrong disambiguation of a word in the utterance in the correct choice

This section provides examples for item 1, 2 and 3.
Figure 4 shows the MT results of an imperative sentence in the test material. The whole dialogue was shown in Figure 1. NICT and Google failed to produce an imperative sentence: NICT's output is ungrammatical because *dake* "only" is postposition that requires an NP that precedes it. Even when *dake* is ignored, the remaining part is a declarative sentence with a null subject[6]. Google produced a grammatical sentence but it's also a declarative sentence with a null subject. Yahoo! failed to translate the nuance of "just" but correctly produced an imperative sentence. The failures of the two SMT systems suggest that imperative sentences are relatively infrequent in their training corpus.
Figure 5 shows a question, in which the correct choice (shown in the square brackets [...]) includes the modal verb 'should' that has both deontic ("need to") and epistemic ("certainly") readings. NICT correctly chose the epistemic reading expressed by Japanese word *darou* ("probably") but Google and Yahoo! wrongly chose the deontic reading and translated it to *hitsuyou-ga-ari-masu* ("need to") and *inakerebaike-masen* ("ought to"). As shown in this example, different kinds of modalities 'overloaded' by a single English modal verb are usually expressed by different Japanese phrases. The low RCAs on Google and Yahoo!'s outputs revealed the wrong choice of the modality types may severly damage the readers' comprehension.
Table 6 shows a question, in which the correct choice (underlined) includes ellipsis, and the translations of that sentence by Google, Yahoo!, and Human-Shuffle. Google

EN: Just use my phone.
Translation by NICT (RCA = 0.06) dake denwa-o riyoshi-masu *only phone*-ACC *use*-POL "only, (someone) uses phone" (ungrammatical)
Translation by Google (RCA = 0.39) chodo watashi-no keitaidenwa-o shiyoshi-tei-masu *just now I*-GEN *cell phone*-ACC *use*-PROG-POL " (someone) is using my cell phone just now"
Translation by Yahoo! (RCA = 0.90) chotto watashi-no denwa-o tsukatte kudasai *a little I*-GEN *phone*-ACC *use* *please* "Use my phone a little, please"

Figure 4: MT results of an imperative sentence in the test material (Question ID: 206)

wrongly supplied "(do not) connect" as the omitted main clause, which does not make sense in the dialogue[7]. Yahoo!'s output includes no such 'supplementation,' but is incorrect and fairly unnatural. The low RCA implies most of the subjects could not infer the original meaning from it.

Translation of the sentences including ellipsis is difficult even for human when the context is not provided. In producing Human-Shuffle translations for such sentences, the translator was instructed to find a correct translation that makes sense in most situation without any supplementation. If it's impossible, the translator supplied a clause assuming a typical situation in which such an utterance is made. We believe this is the best strategy both for human and machine translators that work on individual sentences without considering the context. The Human-Shuffle translation in Table 6 however reveals that the translator could not find a context-independent translation and the assumed context did not match the actual dialogue.

[6]Japanese is a pro-drop language.

[7]On its user interface, Google MT shows which part of an input sentence corresponds to a part of the output. It suggests the phrase "Not this time" is associated to the output *konkai-ha setsuzokushi-masen* ("(We/I) don't connect this time") as a whole in its phrase table.

DIALOGUE

A: We've been driving in circles for an hour. Where are we?
B: I don't know. Let's stop at the next gas station.
A: Yeah. [At least they should have a map.]
B: Why didn't you get one before we left?

System	EN: At least they should have a map.
Google (RCA = 0.06)	Sukunakutomo karera-ha chizu-wo motteiru hitsuyou-ga-ari-masu *at least* *they*-TOP *map*-ACC *have* *need to*-POL "At least they need to have a map"
Yahoo! (RCA = 0.34)	Sukunakutomo karera-ha chizu-wo motte-inakerebaike-masen *at least* *they*-TOP *map*-ACC *have-ought to*-POL "At least they ought to have a map"
NICT (RCA = 0.70)	Sukunakutomo chizu-grai-ha aru-darou *at least* *map-at least*-TOP *exist-probably* "At least there will be a map"

Figure 5: Question involving deontic/epsitemic modality ambiguity and its translations (Question ID: 58)

DIALOGUE

A: Guess which of our students forgot to do the homework last night.
B: I suppose the usual two did, didn't they?
A: [**BLANK**]
B: That's surprising.

OPTIONS

1. No, as usual.
2. Not this time.
3. Yeah, they didn't.
4. Yes, they remembered.

System	EN: Not this time.
Google (RCA = 0.06)	konkai-ha setsuzokusi-masen *this time*-TOP *connect*-NEG "(I / We) do not connect this time"
Yahoo! (RCA = 0.03)	kono-toki-de nai *this time*-COP *not* "(something) is not this time"
Human-Shuffle (RCA = 0.07)	konkai-ha yameteoki-masu *this time*-TOP *decide not to*-POL "I decided not to this time."

Figure 6: Question involving a sentence including ellipsis and its translations (Question ID: 86)

5. Conclusion

We have conducted a task-based MT evaluation involving 795 human participants where dialogue completion questions translated by MT systems were used as the task. Quantitative analysis of the subjects' responses indicated that the following types of frequent translation errors deteriorated the subjects' task performance significantly:

1. Missing words
2. Wrong disambiguations in the lexical choice
3. Literal translations of idioms

Meanwhile, quantitative and qualitative analysis revealed that the following less frequent errors also deteriorated the subjects' performance severely:

1. Incorrect choice of grammatical mood
2. Incorrect choice of the kind of modality
3. Wrong supplementation to omitted phrase
4. Inversion of the sentiment polarity expressed in the source and the translated sentences

Acknowledgments

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

6. Bibliographical References

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Matsuzaki, T., Fujita, A., Todo, N., and H. Arai, N. (2015). Evaluating machine translation systems with second language proficiency tests. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 145–149. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 697–702.