Evaluating Machine Translation Performance In Processing Japanese Ellipsis Strategies

Isabelle Chang

[1] Rutgers University

Evaluating Machine Translation Performance In Processing Japanese Ellipsis Strategies

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants

A total of 14 participants were recruited for this study. Participant age ranged between 19 and 22 years old. All participants were L1 speakers of English, none of which had any formal linguistics training.

Materials

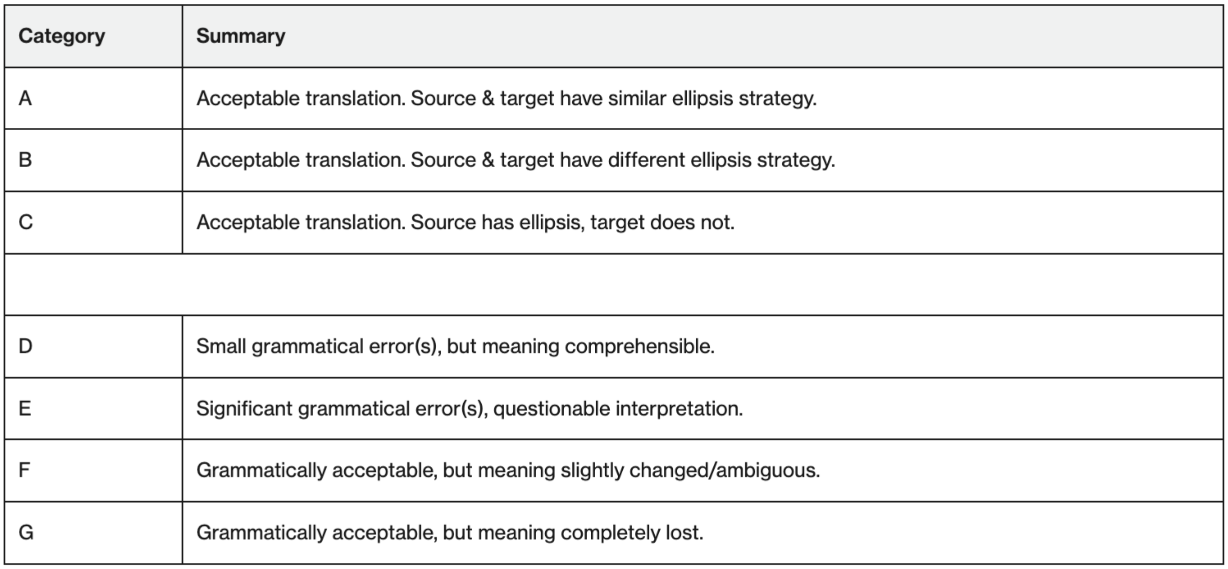
Source Sentences. 15 sentences containing one of three ellipsis strategies present in Japanese were generated or selected from existing literature (five for each ellipsis strategy). The three selected strategies were argument ellipsis, verb-phrase-ellipsis-like constructions, and sluicing. Japanese sentences were generated with the help of two native Japanese speakers. An intended translation was provided for each sentence to serve as a comparison for the machine-generated translations.

Stimuli - Translations. Each of the 15 Japanese sentences was translated using Google, Microsoft Bing, and DeepL Translate (all three services utilize artificial neural networks) for a total of 45 translations.

Acceptability Rating Task

Procedure. Participants were screened for English proficiency prior to completing this task. For the acceptability rating task, participants were presented with each of the 45 machine-generated translations along with an alphabetical scoring scale (A, D, E, F, G) and asked to assign a score to each translation. Additionally, there were two filler questions presented to the participants halfway and three-quarters of the way through the acceptability rating task.

Scoring. The original alphabetical scoring scale was adopted from Khullar 2021:



Category	Summary
A	Acceptable translation. Source & target have similar ellipsis strategy.
B	Acceptable translation. Source & target have different ellipsis strategy.
C	Acceptable translation. Source has ellipsis, target does not.
D	Small grammatical error(s), but meaning comprehensible.
E	Significant grammatical error(s), questionable interpretation.
F	Grammatically acceptable, but meaning slightly changed/ambiguous.
G	Grammatically acceptable, but meaning completely lost.

Figure 1. Alphabetical Scoring Scale

For the present study, B and C scores are removed. This was done in part to account for participants' lack of linguistic background. Additionally, argument ellipsis and VPE-like constructs do not exist in English. As such, any grammatical translation in English would not contain these ellipsis strategies, guaranteeing a B or C ranking for Japanese sentences that contain them. The research questions of this project revolve around the accuracy of translation services—that is, how faithfully translation services are able to retain meaning and grammaticality across across Japanese-English translations. As such, B and C scores were replaced with A scores. Participant-provided alphabetical scores were converted into two sets of numerical scores to quantify the machine translation services' performance on both the meaning and grammaticality of their outputs. The letter-to-number conversion is as follows:

alphabetical_score	meaning_score	grammaticality_score
A	5	5
D	4	2
E	3	1
F	2	4
G	1	3

The original alphabetical scoring scale only evaluates a translation's faithfulness to the source sentence ellipsis strategy (A-C) and meaning (D-G). Adding another scoring scale for grammaticality allows us to evaluate grammaticality and meaning independently of one another.

Results

Summary of Data

14 participants produced 45 acceptability judgments each for a total of 1260 observations (630 for grammaticality and 630 for meaning). Responses to all questions were collected on a Google Form and converted into a .csv file. In the data tidying process, all non-evaluation data were removed. Participant-given alphabetical scores were converted into two separate numerical scores using the system outlined above to yield the following data:

ellipsis_type	mt_service	num_metric	num_score
arg	google	meaning	2
arg	google	grammaticality	4
arg	google	meaning	2
arg	google	grammaticality	4
arg	google	meaning	5
arg	google	grammaticality	5

Descriptive Statistics

Mean scores for meaning and grammaticality by ellipsis type and machine translatin service are listed below:

ellipsis_type	grammaticality	meaning
arg	4.333333	3.780952
sluice	4.076191	4.104762
vpel	3.533333	3.314286

mt_service	grammaticality	meaning
bing	3.766667	3.604762
deepl	4.100000	3.833333
google	4.076191	3.761905

The descriptive statistics suggest that DeepL Translate produces more acceptable translations than Google Translate or Microsoft Bing. Additionally, VPE-like constructs have the lowest average meaning and grammaticality scores, suggesting that this ellipsis type is more difficult to translate than the other two types. The translation services also have higher average grammaticality scores than they do meaning scores, suggesting that services may prioritize the grammaticality over meaning.

Data Analysis

The data were analyzed using a general linear model. Meaning score and grammaticality score were the criteria with ellipsis type and machine translation service as predictors. Main effects of ellipsis type and machine translation service and the ellipsis type by machine translation service interaction were assessed using nested model comparisons. Experiment-wise alpha was set at 0.05.

Results - Meaning

There was a main effect of ellipsis type (F(1) = 16.2938, $p < 0.001$) but not machine translation service on meaning score. There was also main effect of an ellipsis type by machine translation service interaction (F(1) = 7.7691; $p < 0.001$). The model containing the interaction provided the best fit of the data (R2 = 0.085). Although this model is better than the other models tested, it only results in 8.5% of variance in proficiency score values being explained by the interaction between ellipsis type and machine translation service.

Translation scores of sentences originally utilizing VPE-like constructs are affected by the sentence's possession of this ellipsis type. Specifically, sentences with VPE-like constructs yielded translations scoring 0.7 +/-0.24 se points (t=-2.903, $p < 0.004$) lower compared to sentences containing argument ellipsis, which yield and average translation meaning score of 3.74 +/- 0.17 se (t = 21.95, $p < 0.001$).

Results - Grammaticality

There was a main effect of ellipsis type (F(1) = 24.5058, $p < 0.001$) and machine translation service (F(1) = 5.0804, $p < 0.007$). There was also main effect of an ellipsis type by machine translation service interaction (F(1) = 4.1159; $p < 0.003$). The model containing the interaction provided the best fit of the data (R2 = 0.097). Although this model is better than the other models tested, it only results in 9.7% of variance in proficiency score values being explained by the interaction between ellipsis type and machine translation service.

Translation scores of sentences are affects by the source sentence's ellipsis type. Sentences with argument ellipsis have an average translation grammaticality score of 4.26 +/- 0.14 (t = 29.99, $t < 0.001$). Sentences with sluicing score 0.48 +/- 0.20 se (t = -2.4, $p < 0.02$) points lower than sentences with argument ellipsis, and sentences with VPE-like constructions score 1.07 +/- 0.20 se (t = -5.3, $p < 0.001$) lower.

References