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 as follows:

- Argument Ellipsis: Also called nominal ellipsis, this
- Verb Phrase Ellipsis-Like constructions:
- Sluicing: A specific type of clausal ellipsis.

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 (link paper):

(insert scale here)

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:

(insert table with scales here)

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.

Data analysis

We used R (Version 4.2.2; R Core Team, 2022) and the R-packages *papaja* (Version 0.1.1; Aust & Barth, 2022), and *tinylabels* (Version 0.2.3; Barth, 2022) for all our analyses.

Results

Discussion

References

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from https://github.com/crsh/papaja

Barth, M. (2022). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/