Investigating Correlations Between Human Translation and MT Output

Samar A. Almazroei
Kent State University
475 Janik Drive, Kent, OH 44242 | Satterfield H. 109
salmazr1@kent.edu

Haruka Ogawa
Kent State University
475 Janik Drive, Kent, OH 44242 | Satterfield H. 109
hogawa@kent.edu

Devin Gilbert
Kent State University
475 Janik Drive, Kent, OH 44242 | Satterfield H. 109
dgilbe10@kent.edu

Abstract

This study investigates whether there is a correlation between machine translation (MT) and human translation (HT) in terms of word translation entropy (i.e., the variance observed in different translations based on the same source text). Our analysis showed a significant strong correlation in all the three languages we examined: Arabic, Japanese, and Spanish. Furthermore, MT, as well as HT, was found to correlate across languages, although the associations were weaker than the MT-HT correlation in each language.

1 Introduction

This study explores the relationship between the variance in translation output from multiple MT systems and multiple alternative human translations of the same source texts (ST) in three different languages: Arabic, Japanese, and Spanish. Previous studies have reported a correlation between the number of translation options in MT and HT for the same ST words, which leads to the assumption that both MT engines and humans face similar decision-making difficulties within the same language and across different languages (e.g., Carl & Schaeffer 2017, Carl & Báez 2019).

In order to test this hypothesis, the current study first investigates whether the word translation entropy (designated as HTra; see Carl et al. (2016)) of MT output correlates with that of HT in each language. We further investigate to what extent word translation entropy for MT and HT correlates across the three languages. We then conduct qualitative analyses to explore the commonalities and differences among the three languages by comparing the cases where HTra values are high in both MT and HT.

2 Procedure

We used the multiLing texts of the Translation Process Research Database (TPR-DB), which consists of six texts comprising a total of ST 847 tokens and 40 segments. Each text was translated using commercially available MT systems: 12 different systems for Arabic, 13 for Japanese, and 9 for Spanish (for a full list of these systems, see Appendix A).

After obtaining the MT output, the target tokens in each language were aligned componentially to their corresponding English source tokens using Yawat (Germann, 2008). Tokens were aligned on a semantic basis while trying to break phrases down to the smallest units possible, with consistency being key in order for the HTra metric to only reflect output variance and not differences in alignment. For example, if an MT system translates the news story headline "Killer Nurse receives four life sentences" as "*La enfermera del asesino recibe cuatro condenas a cadena perpetua*," 'Killer' would be aligned with '*del asesino*,' 'Nurse' with '*La enfermera*,' 'receives' with '*recibe*,' 'four,' with '*cuatro*,' 'sentences' with '*condenas*,' and 'life' with '*a cadena perpetua*.' The data was then transformed into tables according to TPR-DB conventions. The metric we use in this study (i.e., HTra values) was also calculated according to the same conventions.

3 Results

As shown in Figure 1, results of the Spearman correlation indicated that there is a strong and significant positive association between the HTra of Japanese MT output (HTraJAMT) and that of Japanese HT output (HTraENJA) ($r(845) = .66$, $p < .001$), between Spanish MT output (HTraESMT) and Spanish HT (HTraBML) ($r(845)=.61$, $p<.001$), as well as Arabic MT (HTraARMT) and Arabic HT (HTraAR19) ($r(845)=.62$, $p<.001$). Across the three languages, weak positive correlations were found for MT, and moderate positive correlations for HT (see Figure 1).

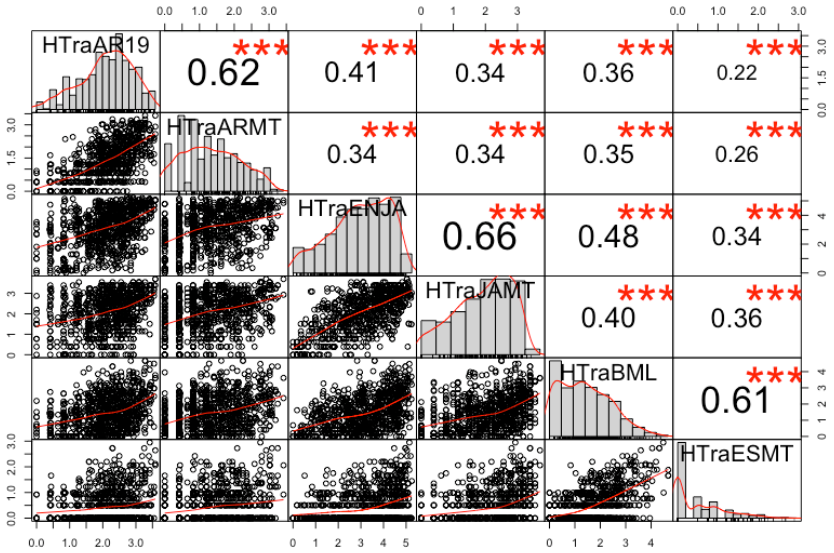

Figure 1: Correlation within and across the three languages

4 Discussion

The correlation between MT and HT was the strongest in Japanese, followed by Arabic and then Spanish. The correlation across languages was moderate for the combination of Arabic and Japanese, and Japanese and Spanish, and weak between Arabic and Spanish. Although the correlations found across languages were weaker than those found within each language, all correlations were still significant (see figure 1).

For qualitative analyses, we ranked the HTra values in each study and examined, for each text in each language, the top 20 tokens and their part of speech (PoS). 51 instances in Arabic and Japanese respectively and 66 in Spanish were found where the HTra values were ranked in the top 20 for both MT and HT. Within these instances, there were only 16 cases where the HTra values were ranked in the top 20 in all the languages. The words "hunter" and "gatherer" in "hunter-gatherer societies" accounted for 6 of these instances. The other instances were mostly idiomatic expressions (i.e., "the extra green mile" and "flaring up") and/or figurative use of verbs (i.e., *hit* as in "Families hit with increase in cost of living" and *flaring up* as in "His withdrawal comes in the wake of fighting flaring up again").

Although all three languages had verb-type tags as the most frequently occurring PoS in their top 20 HTra values, the highest HTra values in the Arabic and Japanese datasets correspond to the 'TO' and 'DT' (determiner) tags, respectively (tags are from the Penn Treebank Project). In the Spanish dataset, however, verb-type tags were the highest and most frequent PoS tags.

5 Remarks

This study reveals intriguing results on the relationship between MT and HT. Further investigations will be conducted to explore whether MT output can be considered as a reliable predictor for human translation effort. In the future, we would like to expand the language variation and examine the commonalities and differences across different languages more qualitatively.

References

Carl, Michael; Schaeffer, Moritz; & Bangalore, Srinivas (2016). The CRITT translation process research database. In *New directions in empirical translation process research* (pp. 13–54). Springer.

Carl, M., & Schaeffer, M. J. (2017). Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. HERMES-Journal of Language and Communication in Business, (56), 43-57.

Carl, M., & Báez, M. C. T. (2019). Machine translation errors and the translation process: a study across different languages. Journal of Specialised Translation, (31), 107-132.

Germann, Ulrich (2008). Yawat: yet another word alignment tool. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, 20–23. Association for Computational Linguistics.

Appendix A. MT Systems Used

Arabic: Amazon Translate, Bing, DayTranslations, Google, Online English Arabic Translator, Prompt Online, Reverso, Systran, Tradukka, Translator.eu, Translatr, and Yandex.

Japanese: Baidu, Bing, Excite, Google, Paralink ImTranslator, Infoseek, MiraiTranslate, Pragma, So-Net, Textra, Weblio, WorldLingo, and Yandex.

Spanish: Amazon Translate, Baidu, Bing, DeepL, Google, Lilt, Pragma, Yarakuzen, and Yandex.