Modelling word translation entropy and syntactic equivalence with machine learning

Bram Vanroy, Orphée De Clercq, Lieve Macken
LT[3], Language and Translation Technology Team
Ghent University
Belgium
firstname.lastname@ugent.be

Previous research suggests that translation product features such as word translation entropy (WTE) and the degree of syntactic equivalence (SE) correlate with cognitive load (Schaeffer et al. (2016)), and Sun (2015), respectively). WTE quantifies the number of translation choices at word level that a translator is confronted with, whereas SE quantifies the syntactic (dis)similarity between a source and target text. In Vanroy et al. (2019), we found that when a source word has multiple possible translations (WTE), a translator may require more cognitive effort to find the suitable translation; and different syntactic structures of the source segment vis-à-vis the proposed target segment may lead to an increased cognitive effort (SE). Consequently, a high average WTE or dissimilar syntactic structures for a given source text and its translation would indicate that a text was difficult to translate.

The current research aims to predict WTE of a source text as well as its SE to a target text without having access to the actual translation products. We do that by training machine learning (ML) systems on a parallel corpus to model these features. We focus on English to Dutch translation, and we use the Dutch Parallel Corpus (Macken et al. (2011); DPC) as our parallel dataset. Unlike the work done in the Translation Process Research Database (Carl et al., 2016) which uses multiple translations of the same text, we calculate a word's translation entropy based on how it has been translated across the whole corpus. We investigate different ML architectures, and features ranging from the sentence to the morphosyntactic level (for the latter, see Tezcan et al. (2017)). The goal is that by only feeding a source sentence into the systems, they can predict that sentence's average WTE and SE.

In addition, we investigate whether we can go one step further and use machine translation (MT) systems as an approximation for human translations for the specific task above. This would mean that we do not need human translations nor ML, and that we can confidently use MT to generate a translation and calculate WTE and SE between the source text and the machine translated target text. To explore the feasibility of this approach, we reuse WTE and SE that were calculated on DPC. Then we translate the source text of that corpus with MT and calculate WTE and SE for these translations. Correlating the WTE and SE values from the human translations and those of the MT version indicates how confidently MT can be used as a proxy for human translations in this task.

This study is carried out in the framework of the PreDicT project[1] (Predicting Difficulty in Translation), which aims to develop a translatability prediction system for English-Dutch that not only automatically assigns a global difficulty score to a given source text, but also identifies the passages in the source text that are difficult to translate.

References

Carl, M., Schaeffer, M. J., and Bangalore, S. (2016). The CRITT translation process research database. In Carl, M., Bangalore, S., and Schaeffer, M. J., editors, *New directions in empirical translation process research*, New frontiers in translation studies, pages 13–54. Springer, Cham, Switzerland.

[1] https://research.flw.ugent.be/en/projects/predict

Macken, L., De Clercq, O., and Paulussen, H. (2011). Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta: Journal des traducteurs*, 56(2):374–390.

Schaeffer, M., Dragsted, B., Hvelplund, K. T., Balling, L. W., and Carl, M. (2016). Word translation entropy: Evidence of early target language activation during reading for translation. In Carl, M., Bangalore, S., and Schaeffer, M., editors, *New directions in empirical translation process research*, pages 183–210. Springer International Publishing, Cham, Switzerland.

Sun, S. (2015). Measuring translation difficulty: Theoretical and methodological considerations. *Across languages and cultures*, 16(1):29–54.

Tezcan, A., Hoste, V., and Macken, L. (2017). A neural network architecture for detecting grammatical errors in statistical machine translation. *The Prague bulletin of mathematical linguistics*, 108(1):133–145.

Vanroy, B., De Clercq, O., and Macken, L. (2019). Correlating process and product data to get an insight into translation difficulty. *Perspectives*, pages 1–18.