Measuring cross-lingual syntactic similarity

Bram Vanroy, Orphée De Clercq, Arda Tezcan, & Lieve Macken

In light of the PreDicT[1] project, we developed cross-lingual syntactic metrics to measure how (dis)similar two sentences are in terms of their syntax (Vanroy, De Clercq, et al., 2021). Whereas we initially created these metrics to investigate difficulties in translation, they can be used for any monolingual or bilingual problem where two related sentences need to be compared.

The syntactic metrics are focused on different linguistic aspects. It is possible to compare the word reordering of one word-aligned sentence with another, or investigate (linguistic) word group reordering. Changes in terms of dependency or part-of-speech label can also be extracted. Finally, aligned syntactic tree edit distance (ASTrED) measures how the dependency structure of the sentences differs while also considering their word alignments.

Within the PreDicT project, these metrics were used to measure the effect of syntactic differences on a translator's total reading time of the source text as a proxy for translation difficulty (Vanroy, Schaeffer, et al., 2021). We found that such diverging syntax has a significant effect on the translation process, indicating that translations that require many syntactic changes impact the translation process significantly and can be empirically considered more difficult.

We have also used the metrics to distinguish the translation habits of different human translators of the same texts. The word order metric, for instance, indicated that some translators translate more literal (less reordering) than others (Vanroy & Macken, 2022). In other work, they have been used to compare the syntactic changes between a literary source text (ST) and a human translation (HT) on the one hand, and ST and neural machine translations (MT) on the other (Webster et al., 2020). They found that MT stays much closer to ST syntax than HT does, who translates more creatively.

Our tool, under the name ASTrED, can be used to automatically parse and align two sentences, and then calculating the metrics mentioned here. Two sentences can be, for instance, a source text and its translation, or a machine translation and a reference translation. It is freely available to use[2] and a demo is available.[3]

References

Daems, J. (2016). *A translation robot for each translator* [PhD thesis]. Ghent University.

Vanroy, B., De Clercq, O., Tezcan, A., Daems, J., & Macken, L. (2021). Metrics of syntactic equivalence to assess translation difficulty. In M. Carl (Ed.), *Explorations in empirical translation process research* (Vol. 3, pp. 259–294). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-69777-8_10

Vanroy, B., & Macken, L. (2022). LeConTra: A Learner Corpus of English-to-Dutch News Translation. *13th Edition of Its Language Resources and Evaluation Conference 2022*, accepted.

Vanroy, B., Schaeffer, M., & Macken, L. (2021). Comparing the Effect of Product-Based Metrics on the Translation Process. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.681945

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

[2] https://github.com/BramVanroy/astred/

[3] https://lt3.ugent.be/astred-demo/

Webster, R., Fonteyne, M., Tezcan, A., Macken, L., & Daems, J. (2020). Gutenberg goes neural: Comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *Informatics*, *7*(3), 32. https://doi.org/10.3390/informatics7030032