FACULTY OF ARTS AND PHILOSOPHY

LT3 – LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

language and translation technology team

Bram Vanroy, Orphée De Clercq and Lieve Macken

USING THE DUTCH PARALLEL CORPUS TO CALCULATE ENGLISH-DUTCH WORD TRANSLATION ENTROPY

A study situated within the PreDicT project*

Goal of PreDicT

Translatability Prediction System

System predicts **difficulty of translating text in language X to target text in Y** as well as **highlight difficult segments**

Input

Source text in language x

Analysis

- **Syntactic** analysis of every sentence, down to morphological level
- **Semantic** analysis
- Recognising **language pair difficulties** using bilingual model

- **Merge** all sentence analyses together
- **Repeat and expand** sentence analysis procedure on larger scale to find issues such as coreference resolution or temporal complexity

Output

Criterion 1 4
Criterion 2 2
Criterion 3 7
Criterion 4 0
Criterion 5 3
Criterion .....
Difficulty score: X

Analysis produces a **global translation difficulty score** for source text X as well as **highlights the segments** in that text that give rise to difficulties

First steps: a pilot study

Aim: correlate translation process data with translation product data

- *Translation process data* (as proxy for cognitive effort):
  - duration (avg. pause ratio, pause dur., production dur.)
  - revision (nr. characters deleted/inserted, nr. edits, non-linear text production)
  - gaze (nr. fixations on source/target text)
- *Translation product data* (difficulty indicators, as shown in literature):
  - nr. errors in translation (Daems, Macken, & Vandepitte, 2013)
  - word translation entropy (Campbell, 2000)
  - amount of (non-)equivalence (Sun, 2015)

Dataset: 690 segments translated by 23 translators, taken from the ROBOT project (Daems, 2016). **Word translation entropy** (WTE) and other features are calculated automatically by CRITT's TPR-DB scripts (Carl et al., 2016). In particular, to calculate WTE the scripts use the translations themselves – which means there is only a very small base corpus.

Findings: all three product features correlate with some process features, especially with the number of times a translator has revised a segment's translation and with the period of pause relative to the segment's total translation time.

Discussion: Calculating word translation entropy on the target text produced by translators is problematic when we need WTE *before* translation has taken place (as is the goal of PreDicT). Therefore, we investigate whether we can generate it *off-line*.

Current study: calculate WTE *off-line*

Aim: To verify whether we can use WTE generated from a parallel corpus:

- we use a larger parallel corpus (DPC, Macken, De Clercq, & Paulussen, 2011) and calculate WTE for all content words
- we re-calculate WTE for the ROBOT data and restrict ourselves to content words
- for both WTEs above, we re-calculate avg. WTE per segment in the ROBOT dataset
- we calculate correlations between process data and each version of WTE
- we compare the correlations to see if using WTE gathered from a large corpus is feasible

Findings: using WTE based on the small ROBOT corpus and WTE based on DPC yields the same results when correlating with particular revision information from the ROBOT process data, namely the nr. edits. Literature proposed that this was a good intermediary showing cognitive effort.

Conclusion: with respect to correlating to cognitive effort indicators (and, thus, being a difficulty indicator), WTE can be modelled off-line by using large corpora as its base. This allows us to add WTE to our pipeline without the need of any translation.

References

Campbell, S. (2000). Choice Network Analysis in Translation Research. In M. Olohan (Ed.), *Intercultural Faultlines: Research Models in Translation Studies: Textual and Cognitive Aspects* (pp. 29–42). Manchester, UK: St. Jerome.

Carl, M., Schaeffer, M. J., & Bangalore, S. (2016). The CRITT Translation Process Research Database. In M. Carl, S. Bangalore, & M. J. Schaeffer (Eds.), *New Directions in Empirical Translation Process Research* (pp. 13–54). Cham, Switzerland: Springer International Publishing.

Daems, J. (2016). *A Translation Robot for each Translator* (PhD Thesis). Ghent University, Ghent, Belgium.

Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the Sum of its Parts: A Two- Step Approach for the Identification of Translation Problems and Translation Quality Assessment for HT and MT+PE. In S. O'Brien, M. Simard, & S. Lucia (Eds.), *Proceedings of MTS 2013 Workshop on Post-Editing Technology and Practice* (pp. 63–71). Nice, France.

Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta: Journal Des Traducteurs*, *56*(2), 374–390.

Sun, S. (2015). Measuring Translation Difficulty: Theoretical and Methodological Considerations. *Across Languages and Cultures*, *16* (1), 29–54.

*** PreDict project** (2017-2021)
Predicting Difficulty in Translation
http://research.flw.ugent.be/en/projects/predict

Contact

Bram.Vanroy@UGent.be
www.lt3.ugent.be/people/bram-vanroy/
/bramvanroy

GHENT UNIVERSITY