language and translation technology team

LITERARY MT UNDER THE MAGNIFYING GLASS: ASSESSING THE QUALITY OF AN NMT-TRANSLATED DETECTIVE NOVEL ON DOCUMENT LEVEL

MARGOT FONTEYNE, ARDA TEZCAN, LIEVE MACKEN

Introduction

Several studies have demonstrated that translation quality has improved enormously since the emergence of NMT[1, 2, 3]

But: usually sentence-level evaluations of general text types

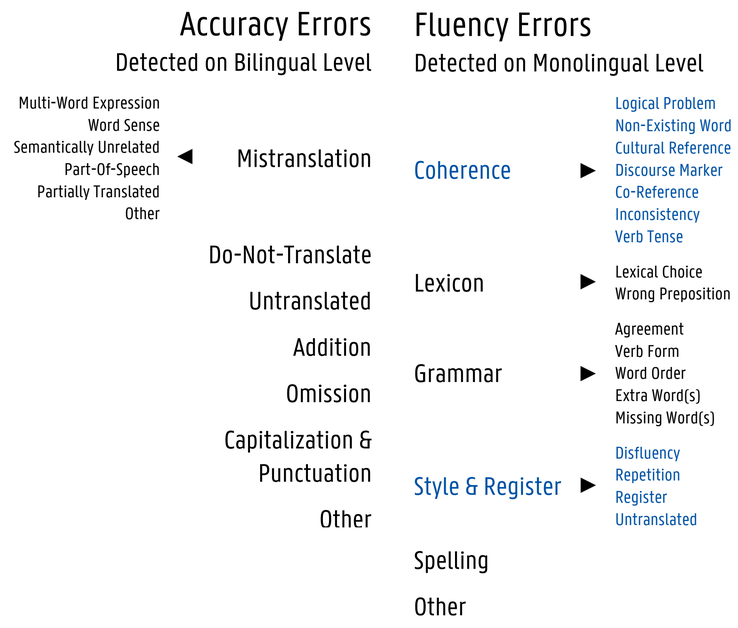
- What happens when we carry out the evaluation on machine translations of more creative text types, such as literature?
- What happens when we carry out the evaluation on document level?

Text Selection

- Detective novel by Agatha Christie
- 58,110 words – 5,276 sentences
- Motivation: eye-tracking data for human translation available (future work: gather corresponding data for MT)
- Translated from English into Dutch by Google's NMT system

Fine-Grained Error Annotation

- Entire novel annotated by one annotator
  First chapter also independently annotated by a second annotator
- WebAnno annotation tool
- Errors classified within the hierarchical SCATE MT error taxonomy (3 levels), adapted to literary NMT on document level[4]
- Two-step annotation approach: first fluency, then accuracy
- Annotations spans are allowed to overlap



Extended SCATE MT error taxonomy with document-level features

Step 1: annotating all fluency errors without access to the source

Step 2: annotating & linking all accuracy errors in source & target

Inter-Annotator Agreement (IAA)

Do the annotators agree on their annotations on the first chapter?

(Dis)agreement on error detection

- Low IAA when calculated on word level (only 56% of the annotated words were annotated by both annotators)
- Decent IAA when calculated on annotation level (73% of the annotations overlap with an annotation by the other annotator)
- Explanation: tendency to disagree on the length of the error span

(Dis)agreement on error categorization

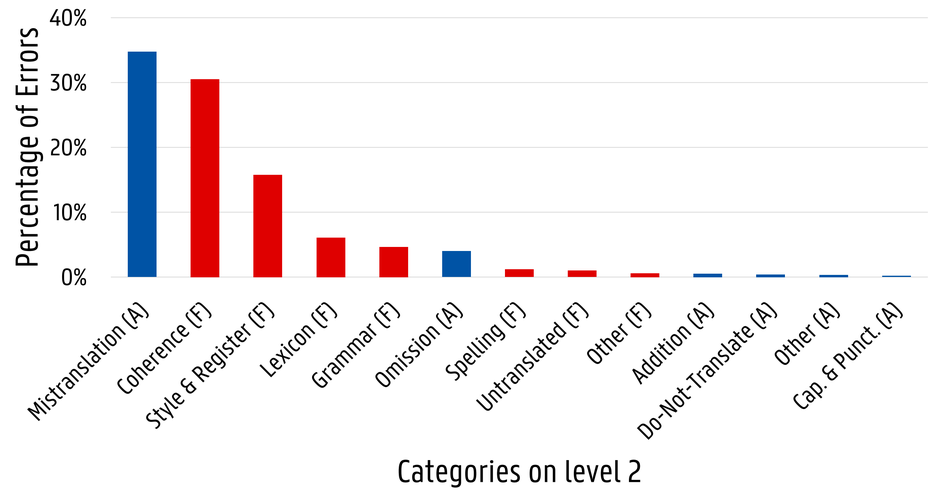
- 99% of the overlapping annotations agree on level 1 of the taxonomy, 90% on level 2, and 73% on level 3
- Most pronounced disagreement on level 2: 'coherence' vs. 'style & register' (5% of the annotation pairs agreeing on level 1)

Error Analysis

Overall Quality

- 44% of the sentences do not contain any errors
- On average, the erroneous source sentences are longer than the correct ones (17.11 vs. 9.56 words)
- Performance decreases from a source sentence length of 10 words onwards

Error Distribution

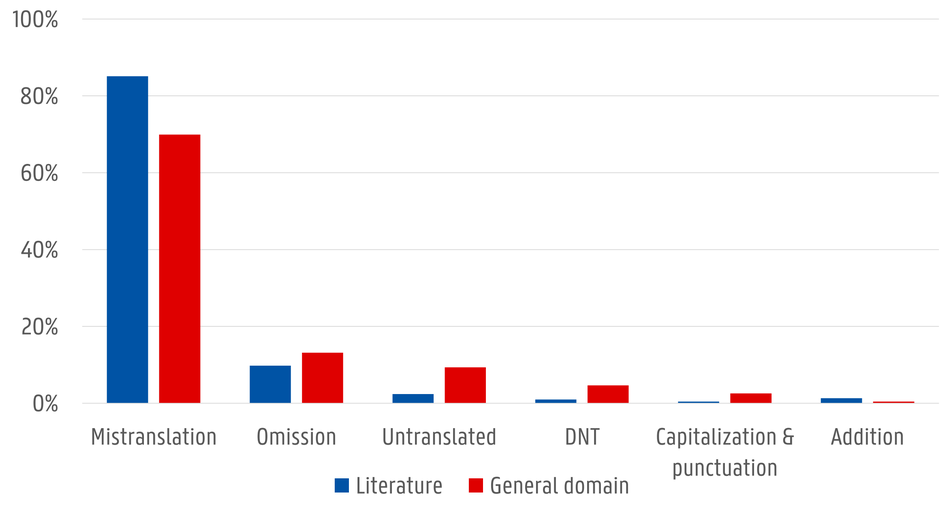


Co-Occurring Fluency & Accuracy Errors

Hypothesis: if accuracy errors tend to cause fluency errors, they will overlap regularly

- 27% of the fluency and 39% of the accuracy error spans fully overlap with the span of an error belonging to the other category
- 84% of those overlaps: 'coherence' x 'mistranslation'

Accuracy Comparison with General Domain[5]



Conclusions

- Literary MT seems to be more promising for English-Dutch (44% correct sentences) than for English-Slovene (0%)[6], English-Russian (17%)[7] & German-English (33%)[7]
- Main problems for literary MT are 'mistranslation', 'coherence' & 'style & register'
- Mistranslations are the biggest accuracy issue for general-domain MT as well.
- Fluency & accuracy errors overlap regularly, especially 'coherence' & 'mistranslation' errors

GHENT UNIVERSITY

[1] Bentivogli, L., Bisazza, A., Cettolo, M. and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, November 1-5. Association for Computational Linguistics, pp. 257-267.
[2] Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T. and Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108, pp. 159-170.
[3] Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, April 3-7. Association for Computational Linguistics, pp. 1063-1073.
[4] Tezcan, A., Daems, J. and Macken, L. (2019). When a 'sport' is a person and other issues for NMT of novels. In *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, 19 August. European Association for Machine Translation, pp. 40-49.
[5] Van Brussel, L., Tezcan, A. and Macken, L. (2018). A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12. European Language Resources Association (ELRA), pp. 3799-3804.
[6] Kuzman, T., Vintar, Š. and Arčan, M. (2019). Neural machine translation of literary texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, 19 August. European Association for Machine Translation, pp. 1-9.
[7] Matusov, E. (2019). The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, 19 August. European Association for Machine Translation, pp. 10-19