With a Little Help from the Authors: Reproducing Human Evaluation of an MT Error Detector

Ondřej Plátek, Mateusz Lango, Ondřej Dušek

September 7, 2023

ÚFAL

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

The reproduced paper

- Jannis Vamvas and Rico Sennrich, *As Little as Possible, as Much as Necessary: Detecting Over- and Undertranslations with Contrastive Conditioning*, ACL 2022
- What are over/under-translations?

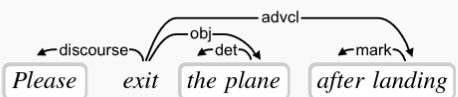
Estamos en Varna → We are in Varna, Bulgaria.

- Overview of the approach

1 Translate

X = *Please exit the plane after landing.*
Y = *Bitte verlassen Sie das Flugzeug.*

2 Extract constituents



3 Score conditioned on partial sequences

Score(Y | *Please exit the plane after landing.*) = 0.34
Score(Y | *~~Please~~ exit the plane after landing.*) = 0.14
Score(Y | *Please exit ~~the plane~~ after landing.*) = 0.20
Score(Y | *Please exit the plane ~~after landing~~.*) = **0.72**

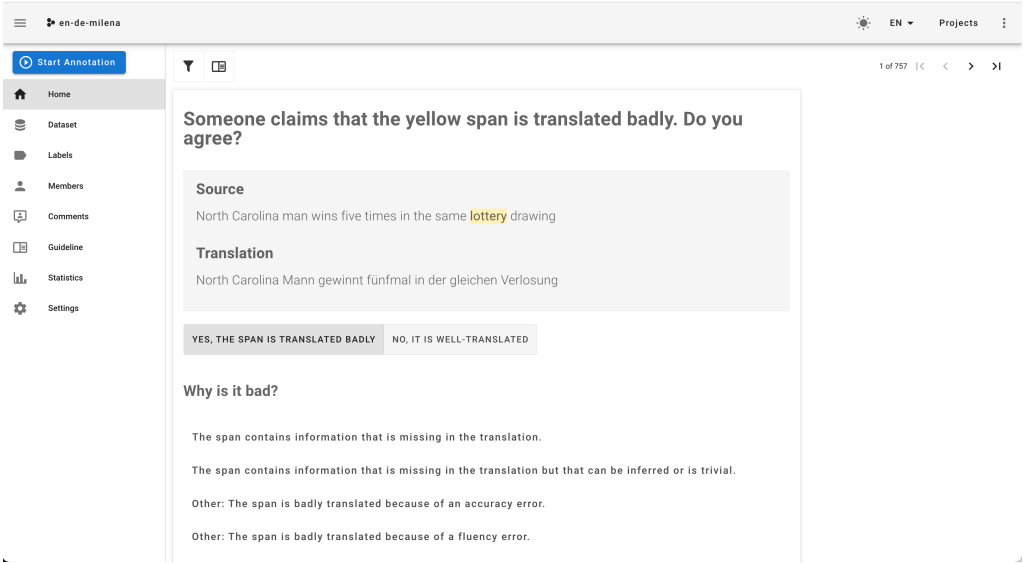
4 Infer error spans

Please exit the plane after landing .

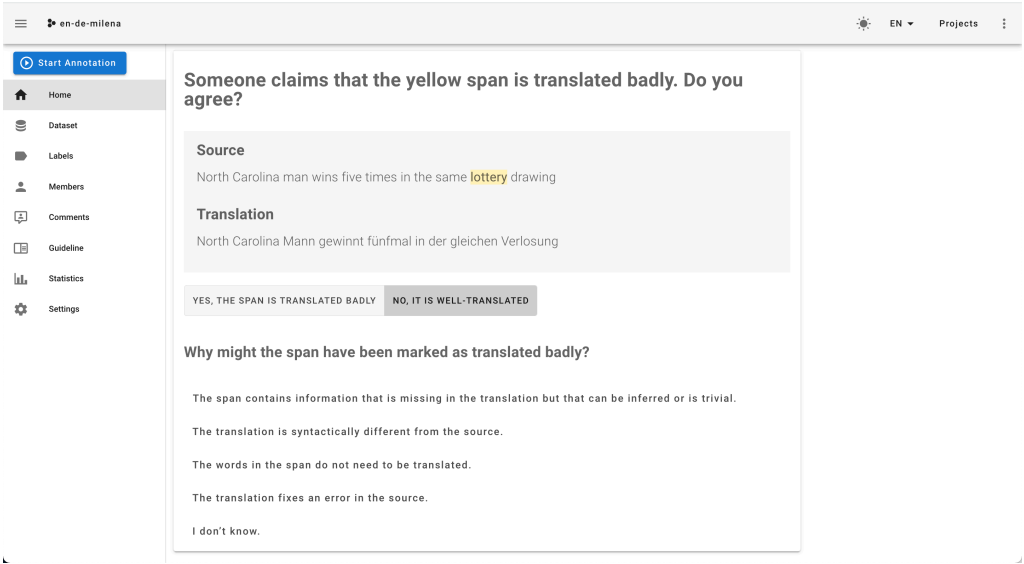
The reproduced paper – human evaluation

- Goal of the human evaluation: assess correctness of detected over/under-translations
- Two types of results:
  - precision of the method in indicating translation errors
    - in general
    - over vs. under-translations
  - fine-grained analysis - reasons behind indicating an error
    - lack of fluency, syntactic differences, ...
- English-German & English-Chinese evaluated
- For each language pair, 2 linguists annotated 700 detected over/under-translations

The reproduced paper – human evaluation



The reproduced paper – human evaluation



Our reproduction

- Our reproduction: English-German only
- 2 annotators but from German universities (original: Swiss).
- Same annotation guidelines
- Same predictions annotated, but presented in different order
- Issues:
  - Doccano framework extension impossible to run with current version
    - Authors provided a Docker image
  - We found a minor bug in the data aggregation script

Results - coarse-grained precision

		Original	95% CI	Reproduction	CV*
Target (over)	Addition errors	2.3	(1.38; 3.71)	1.95	16.42
	Any errors	7.4	(5.66; 9.68)	6.77	8.86
Source (under)	Omission errors	36.3	(32.57; 40.18)	* 14.23	19.56
	Any errors	39.4	(35.61; 43.34)	* 22.09	15.34

- Precision of detected over-translations is slightly lower (not statistically significant)
- Precision for under-translations is significantly lower

Results - fine-grained analysis

- GOF statistical tests: verify deviation of reproduced vs. original coarse-grained results

		χ^2	**p-value**	V
Overtrans.	good trans.	355.77	$<$0.0001	0.50
	bad trans.	* 201.88	$<$0.0001	0.71
Undertrans.	good trans.	596.99	$<$0.0001	0.57
	bad trans.	* 15.8	0.0016	0.34

- All Cramer's V values are $> 0.29 \sim$ large data distribution discrepancy (Cohen, 1988)
- Repr.: "I don't know" chosen as label $4\times$ more often than original study
- Repr.: translation correct & trivial information missing – 103 counts (vs. 25 orig.)
- Repr.: translation incorrect & trivial information missing – 7 counts (vs. 107 orig.)

Results – inner-annotator agreements

- Krippendorff's alpha coefficient (α) between reproduction & original:
- Coarse-grained analysis

	α	**%Ident.**
Overtranslation	0.6976	0.9558
Undertranslation	0.3762	0.7266
Joint	0.5109	0.8475

- Fine-grained analysis

		α	**%Ident.**
Overtranslation	Good translation	0.2238	0.5059
	Bad translation	0.1982	0.4687
	Joint	0.2607	0.5033
Undertranslation	Good translation	0.1427	0.3365
	Bad translation	0.1994	0.4468
	Joint	0.2084	0.3621
Joint		0.2664	0.4366

Conclusions

The authors of original study draw the following conclusions:

- Precision is higher for undertranslations, but still low for overtranslations
- Many highlighted spans are translation errors, but not over/undertranslations
- Syntactic differences contribute to the false positives for overtranslations

These conclusions are *confirmed* in our reproduction, but the observed effect sizes were considerably lower:

- Precision difference: was 12.28% instead of 34%
- Syntactic differences given ca. 40% less frequently as reason for false positives

Takeaways

- We successfully reproduced the human evaluation of (Vamvas and Sennrich, 2022)
- Despite high-quality documentation, availability of annotation guidelines, etc., assistance from the authors of the original study was necessary and essential for the reproduction
- To ensure reproducibility, annotation interfaces should be well documented and easy to run (e.g. provided as Docker images)
- Coarse-grained results much more consistent with the original study, supporting the experiment design with a very limited number of possible responses

Thank you!

Paper:
https://arxiv.org/abs/2308.06527

Github:
bit.ly/github-reprohum

Ondřej Plátek
@oplatk

Ondřej Dušek
@tuetschek

This research was supported by Charles University projects GAUK 40222 and SVV 260575 and by the European Research Council (Grant agreement No. 101039303 NG-NLG). It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).