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

Ondřej Plátek, Mateusz Lango and **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{oplatek,lango,odusek}@ufal.mff.cuni.cz

Abstract

This work presents our efforts to reproduce the results of the human evaluation experiment presented in the paper of Vamvas and Sennrich (2022), which evaluated an automatic system detecting over- and undertranslations (translations containing more or less information than the original) in machine translation (MT) outputs. Despite the high quality of the documentation and code provided by the authors, we discuss some problems we found in reproducing the exact experimental setup and offer recommendations for improving reproducibility. Our replicated results generally confirm the conclusions of the original study, but in some cases statistically significant differences were observed, suggesting a high variability of human annotation.

1 Introduction

Reproducibility of experimental results is a fundamental principle of scientific research that ensures the validity, credibility, and reliability of scientific findings. The NLP research community is increasingly interested in reproducibility, which leads to the organization of shared tasks (Belz et al., 2021, 2022b; Branco et al., 2020), the formulation of reproducibility guidelines (Pineau et al., 2021), and so on. However, most previous efforts are limited to the reproducibility of automatic measures, and reproducibility of human evaluation has received less attention (Belz et al., 2023). The ReproHum project,[1] which this paper is a part of, aims to improve this situation.

In this paper, we describe our efforts to reproduce the results of the human evaluation experiment conducted in (Vamvas and Sennrich, 2022) to evaluate the performance of their over- and undertranslation detection method for machine translation (MT). More specifically, the method detects

phrases in the source texts whose meaning is not reflected in an MT output, or phrases in the MT output that are not supported by the source (see details in Sec. 2). The human annotators evaluated the detection accuracy and provided additional reasons for their evaluation by choosing from a list (see Figure 1). The original experiment was run for English-Chinese and English-German translation pairs, but our reproducibility study is limited to the English-German pair due to the availability of skilled human annotators.

Despite the precise description of the experiment in the original paper, we still encountered difficulties in running the study (see Section 4), and were only able to finish it successfully thanks to the strong support of the original authors. Our results overall support the main claims of Vamvas and Sennrich (2022)'s original paper, but we still found some discrepancies, despite using the same data, interface and guidelines for the annotation (see Section 5).

Our collected annotation outputs, reproduction code, and the filled HEDS sheet (Shimorina and Belz, 2022) for the reproduction study are available on Github.[2]

2 Original Experiment

The original paper (Vamvas and Sennrich, 2022) proposed an automatic method for detecting coverage errors in the output of MT systems. Coverage errors include *undertranslations*, i.e. the omission of important source content in the MT system output, and *overtranslations*, i.e. the addition of superfluous words to the translation that may not be supported by the source.

The method uses contrastive conditioning (Vamvas and Sennrich, 2021) and finds coverage errors

---

by iteratively computing the probability of the generated translation with an MT system conditioned on an incomplete text source. If the probability of the generated translation increases when a particular phrase is deleted from the source, the method takes this as an indicator that the deleted phrase is not adequately reflected in the translation and treats it as an omission (undertranslation). Similarly, by reversing the source with the target, the method also detects overtranslations. To summarize, for a given input-translation pair, the output of the method is the type of problem detected (over- or undertranslation) and a phrase that has been omitted from the translation (in the case of undertranslation), or that is superfluous (in the case of overtranslation).

The corresponding human evaluation aimed to analyze in detail the predicted problematic text spans and assess their correctness. Human annotators were presented with a source sentence, the generated MT translation, and a highlighted passage. The annotator's task was to decide whether the highlighted passage was correctly translated and later to select additional feedback for fine-grained analysis from a given single-choice list. If the annotator confirmed that the highlighted span was incorrectly translated, they were asked to specify the type of error (e.g., fluency error, accuracy error, addition/omission of non-trivial information, etc). On the other hand, if the annotator considered the span to be correctly translated, they were asked to give a possible reason why one could think that it was translated incorrectly (e.g., syntactic differences, adding/removing trivial/obvious information). The full list of possible reasons can be seen in Figure 1.

The manual evaluation was carried out by two linguists who were provided with a two-page document containing annotation guidelines. The guidelines included the task description, instructions on using the annotation interface, and examples of three incorrect and three correct translations. Each annotator responded to approximately 700 randomly selected examples.

3 Differences in Our Reproduction Study

We aimed to conduct the reproduction as close as possible to the original study. We worked on the same set of system outputs, with the identical annotation interface and instructions.[3] However, there

---

were some differences with respect to annotator hiring and to splitting the annotation between them.

Annotator hiring We hired two annotators who were university students and native speakers of German with high proficiency in English, same as in the original study. We used contacts arranged through ReproHum organizers at two German universities (Bielefeld and Munich), which means that the students spoke a different variety of German from the original study, which was conducted in Zurich, i.e., with Swiss German speakers. In addition, one of our annotators was from a different study field (public health) than the original study's annotators (NLP). The reason is that we could not find interested NLP students at the time of hiring. Each annotator was paid €180 for the approximate 10 hours of work. The €18 hourly wage differs from the original study (which reports ca. $30) but is in line with ReproHum recommendations (150% of German minimum wage).

Data split for annotation We used the same input data for the human evaluation, i.e., the same outputs of the machine translation (MT) system, together with error annotations. Similarly to the original study, the sentences with annotated errors were split randomly into two parts, for one annotator each. However, the original authors prepared for us a different random split of data since the reproduction of the study should not depend on how the original data are split among the annotators.

4 Implementation Issues

We found two implementation problems while running the study: one with setting up the annotation interface and the other with the script computing final statistics.

Annotation interface The authors of the original study used the popular open-source annotation software Doccano (Nakayama et al., 2018), which was customized to implement the interface required for their human evaluation experiment. The original open-source software has been updated over time, making the authors' customization incompatible with the toolkit. Even after downgrading Doccano to the version used by the authors, some of the dependencies were found to be no longer available. Our attempts to use newer versions of dependencies and/or Doccano were unsuccessful.

Finally, with the permission of the ReproHum organisers, we contacted the authors, who fortu-
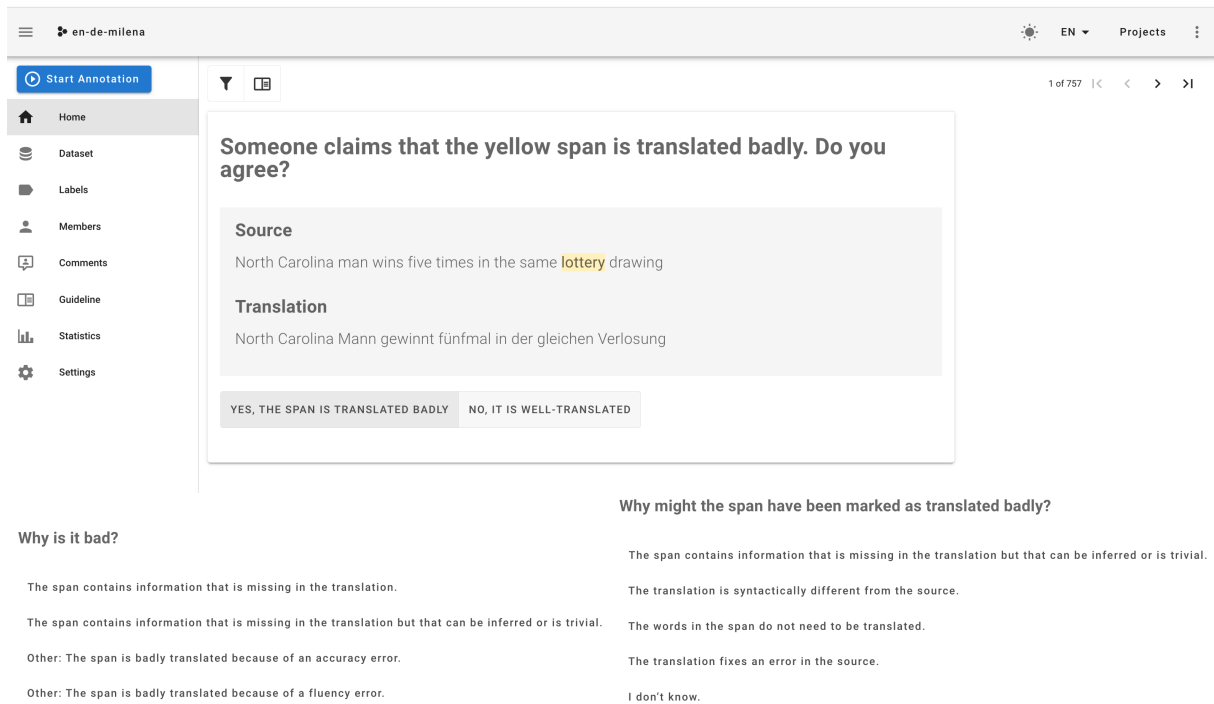
Figure 1: Screenshots of the Annotation Interface where the annotator needs to: (1) Top image; Select whether the source sentence is well translated. (2) Bottom left image; In the case of a bad translation, indicate the type of error. (3) Bottom right image; In the case of a correct translation, hypothesize why it was marked as an error.

nately kept the annotation interface in an easy-to-distribute form of a Docker snapshot (Merkel, 2014). With the Docker image provided, we were able to run the reproduction with the identical annotation interface.

Statistics computation To compute the necessary annotation statistics, we used the evaluation script provided by Vamvas and Sennrich (2022). During our data analysis, we noticed that the script did not correctly handle examples with multiple spans annotated within the same sentence. For such examples, the last annotation analysed by the script would override the previous ones, resulting in some annotations being unintentionally removed. We fixed this bug and ran the analysis with the original and corrected script on the annotation data from both the original and repeated study (see Table 1).

5 Results

A side-by-side presentation of all the results can be found in Table 1. We report results calculated using both the original script (Original, Reproduced) and the corrected script detailed in Section 4 (Original_v2, Reproduced_v2). The use of the corrected script increased the number of examples by 54 for the original results and by 56 for the reproduced results, which represents less than 4% of the total data size. In order to maintain a better reference to the results reported in the original paper, and since the difference is not large, the following discussion will analyse the results computed with the original script.

In both experiments, each annotator provided feedback on approximately 700 examples, which were similarly divided into examples of over- and under-translation. The inter-annotator agreement, measured with Cohen's κ, was higher in our reproduction for the simple evaluation of the correctness of highlighted spans, but the results were within the constructed 95% confidence interval for the original value. According to McHugh (2012), these κ values should be interpreted as weak/moderate agreement. On the other hand, the inter-annotator agreement for fine-grained responses was statistically significantly lower in our repeated study, but again, both values from the original and the repeated study should be interpreted as "minimal agreement" (McHugh, 2012).

In the original paper, the results of the experiments are reported in terms of word-level precision of the highlighted spans. The comparison of the original precision values with those obtained in our reproduction can be observed in Table 2. In addition, for each original value, we computed

	Orig	Repro	Orig_v2	Repro_v2
Basic statistics				
Annotator 1 no. of overtranslation examples	372	372	390	389
Annotator 1 no. of undertranslation examples	344	348	354	361
Annotator 1 total no. of examples	716	720	744	750
Annotator 2 no. of overtranslation examples	348	352	362	366
Annotator 2 no. of undertranslation examples	351	350	363	362
Annotator 2 total no. of examples	699	702	725	728
Inter-annotator agreement				
Number of overlapping samples	466	471	474	479
Cohen's κ for trans. correctness eval.	0.5343	0.6044	0.5593	0.6313
Cohen's κ for fine-grained answers	0.3186	0.2420	0.3328	0.2536
Precision of spans indicated by the method as incorrect				
Incorrectly translated spans detected as overtanslations	49	45	54	48
Correctly translated spans detected as overtanslations	611	619	638	647
Incorrectly translated spans detected as undertranslations	240	135	250	143
Correctly translated spans detected as undertranslations	369	476	379	491
Fine-grained analysis for overtranslations				
True overtranslations - addition of trivial or inferable information	10	2	10	2
True overtranslations - addition of unsupported information	5	11	5	12
True errors - accuracy errors	28	24	32	24
True errors - fluency errors	6	8	7	10
False errors - addition of redundant but fluent info.	113	120	117	124
False errors - addition of supported information	19	30	20	30
False errors - a syntactic difference	428	254	449	263
False errors - unknown reason	51	215	52	230
Fine-grained analysis for undertranslations				
True undertranslations - lack of important information	114	80	120	86
True undertranslations - lack of redundant information	107	7	110	7
True errors - accuracy errors	16	35	17	37
True errors - fluency errors	3	13	3	13
False errors - fluency errors in the source	72	107	72	110
False errors - addition of redundant but fluent info.	25	103	25	104
False errors - a syntactic difference	249	174	257	178
False errors - unknown reason	23	92	25	99

Table 1: The summary of the raw results obtained in the original and reproduced study. The columns with "v2" suffix are computed with the fixed evaluation script (see Sec. 4). Note that annotators for Original and Reproduced studies differ: Annotator 1 is a different different person for the Original(_v2) and Reproduced(_v2) studies. Similarly for Annotator 2.

		Original	95% CI	Reproduction	CV*
Target	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	Omission errors	36.3	(32.57; 40.18)	* 14.23	19.56
	Any errors	39.4	(35.61; 43.34)	* 22.09	15.34

Table 2: Word-level precision (%) of the spans that were highlighted by the method (Vamvas and Sennrich, 2022, Table 2) in the original study and in our reproduction, together with 95% confidence intervals constructed for the original values (95% CI) and the small-sample coefficient of variation (CV*). Reproduced results that do not fall within the CI are marked with an asterisk.

		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

Table 3: The results of goodness-of-fit (GOF) tests of human answers in fine-grained analysis (types of error) in the original and our reproduced study. The effect size is measured with Cramér's V for GOF. For cases marked with an asterisk, the conditions to use χ^2 approximation were not met, thus the test statistics were estimated with Monte Carlo simulation (10k samples).

a 95% confidence interval using Wilson's score method (Wilson, 1927). The precision values obtained in our reproduction are generally lower than those reported in the original study. However, the differences in precision for the overtranslation spans are still within the confidence intervals. In contrast, the differences for under-translation are substantial, as the reproduced precision values are about 44-46% lower. This difference is also statistically significant at the significance level $\alpha = 5\%$.

To compare the results of the fine-grained analysis, we performed χ^2 goodness-of-fit tests between the answers provided by the original annotators and those provided by us, as well as calculating Cramér's V for goodness-of-fit. The results are presented in Table 3. We were able to reject the null hypothesis that the reproduced fine-grained responses follow the distribution of the original responses with low p-values for all four sets of results. All obtained values of Cramér's V exceed the 0.29 threshold suggested by Cohen (1988) as an indicator of a large discrepancy between the data distributions.

The differences are visible even to the naked eye, as our annotators selected unknown reasons for highlighting a correctly translated text span about four times more often than in the original study.

In the case of undertranslations, the annotators in the original study chose much more often that the translation is incorrect, but the missing information can be inferred or is trivial (107 vs. 7 counts). On the other hand, our annotators were much more likely to consider that the translation was correct but could be considered inaccurate because some trivial or easily inferable information was missing (25 vs. 103 counts). Our annotators also found more spans highlighted as under-translations as reasons for accuracy or fluency errors in the trans-

lation. These larger differences for examples of undertranslation may also indicate that this variant of the evaluation task is more difficult. Note that the annotators have a highlighted text span in the source text, but still have to answer questions about the final translation without any word/phrase alignment information.

6 Quantifying Reproducibility

Following the guidelines of the ReproHum shared task (Belz et al., 2023, Sect. A5), we identify reproduction targets in the following categories:

- Type I – numerical scores:

  - the precision of text spans labeled as over-/undertranslations to truly contain over-/undertranslation errors
  - the precision of text spans labeled as over-/undertranslations to contain some translation errors

- Type II – sets of numerical values:

  - the set of precision results for examples marked as overtranslations
  - the set of precision results for examples marked as undertranslations

- Type III – categorical labels attached to text spans:

  - Sets of spans annotated with the correct/incorrect translation label, separately for over- and under-translations.
  - Sets of fine-grained reasons given by annotators for marking a span as incorrect, separately for over- and under-translations and for correctly/incorrectly detected spans.

Type I For the numerical results, we followed the quantified reproducibility assessment by Belz et al. (2022a), which involves calculating the small sample coefficient of variation (CV*) as a measure of the degree of reproducibility. The results are given in the last column of Table 2. Three out of four CV* values are in the 15-20 range. Only for the precision of detecting a translation error in the text span marked as overtanslation, the CV* value is significantly lower (8.86).

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

Table 4: Krippendorff's alpha coefficient (α) for assessment of bad/good translation using combined annotations for both original and replication studies. %Ident. is the percentage of identical answers between the original and replicated annotation. (The annotations of both annotators were combined into one list. For examples where both annotators provided an answer, half of the answers were taken from each annotator.)

		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

Table 5: Krippendorff's alpha coefficient (α) for fine-grained analysis using the combined annotations (see comment in Table 4). %Ident. is the percentage of identical answers between the original and replicated annotation.

Type II results are usually evaluated with Pearson's correlation (Huidrom et al., 2022), but there is little point in calculating it here. It is equal to 1 for both over- and under-translations, while the standard statistical test for correlation fails to reject the null hypothesis that the true correlation is 0.

Type III Finally, the reproducibility of categorical labels was assessed with Krippendorff's alpha. Since the aim of this analysis is not to measure the agreement between all four annotators (2 from the original study and 2 from the replication) but rather to measure the reproducibility, for the purpose of computing Krippendorff's alpha the annotations obtained from each pair of annotators were combined into one set. For the overlapping examples i.e. examples annotated by both annotators, the response of a randomly chosen annotator was retained.

The values of Krippendorff's alpha together with the percentage of identical evaluations in both the original and reproduced study for coarse-grained annotations (i.e. correct/incorrect translation label) are given in Tab. 4. The value for overtranslation is significantly higher than for undertranslations and is above the $^2/_3$ threshold suggested by Krippen-

dorff (2004) as the lowest limit to consider a good agreement between the raters. The high agreement between the original annotation and the reproduced one can also be observed in terms of the proportion of identical answers – almost 85% for the whole dataset.

Similar results for fine-grained annotations are provided in Tab. 5. All reported values of Krippendorff's alpha are relatively low, and the proportion of examples that are evaluated identically in both studies is below 51%. However, the ratio for overtranslation spans marked as correct translations is significantly higher than the same ratio for undertranslations (roughly 16 percentage points).

7 Findings

Based on the manual evaluation, the authors of the original paper present several findings/conclusions:

- Precision is higher for undertranslations, but still low for overtranslations

- Many of the highlighted spans are translation errors, but not over/undertranslations

- Fine-grained analysis suggests that syntactic differences contribute to the false positives for overtranslations.

All of the above findings roughly correspond to the results of our reproduced experiment. The precision for undertranslations is also higher than for overtranslations, but the difference between the two was considerably smaller in our experiment. For example, the difference in precision for true coverage errors is 34.0 percentage points in the original study but only 12.3 in ours. Similarly, our fine-grained analysis confirms that syntactic differences contribute to false positives, but they are reported about 40% less frequently than in the original study. However, as mentioned above, this is partly due to the fact that our annotators were much more likely to select the "unknown reason" response.

8 Conclusion

We carefully repeated a human evaluation study of paper (Vamvas and Sennrich, 2022). Despite the high-quality documentation and well-organized code provided by the authors, we encountered several problems that were difficult to overcome. In particular, we would not have been able to run the annotation interface and repeat the study without

the authors' help. We also noticed a minor error in their evaluation script and suggested a modification. The reproduction process made us realise that it is almost impossible to publish a fully reproducible paper without actually trying to reproduce it end-to-end. We also advocate distributing annotation interfaces in the form of a Docker image containing all the dependencies.

Our overall results agree with the high-level conclusions made in the original paper. The reproduction of results of the coarse-grained analysis was successful for overtranslations, but the results for undertranslations were significantly lower than the reported in the original study. The results of the fine-grained analysis were even more difficult to reproduce – we observed significant differences in all analyzed groups of answers. This may suggest that when designing experiments with human judgments, setups with a very limited number of possible answers (especially binary questions) are easier to replicate and should be prioritized over more complex setups whenever possible.

Acknowledgments

References

Anya Belz, Maja Popovic, and Simon Mille. 2022a. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022b. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Huiyuan Lai, Chris van der Lee, Emiel van Miltenburg, Yiru Li, Saad Mahamood, Margot Mieskes, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Pablo Mosteiro Romero, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Rudali Huidrom, Ondřej Dušek, Zdeněk Kasner, Thiago Castro Ferreira, and Anya Belz. 2022. Two reproductions of a human-assessed comparative evaluation of a semantic error detection system. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 52–61, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage publications.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Dirk Merkel. 2014. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle.

2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *The Journal of Machine Learning Research*, 22(1):7459–7478.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.