Robustness of Multi-Source MT to Transcription Errors

Dominik Macháček[1,2] and **Peter Polák**[1] and **Ondřej Bojar**[1] and **Raj Dabre**[2]

Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics[1]

National Institute of Information and Communications Technology, Kyoto, Japan[2]
{machacek,polak,bojar}@ufal.mff.cuni.cz, raj.dabre@nict.go.jp

Abstract

Automatic speech translation is sensitive to speech recognition errors, but in a multilingual scenario, the same content may be available in various languages via simultaneous interpreting, dubbing or subtitling. In this paper, we hypothesize that leveraging multiple sources will improve translation quality if the sources complement one another in terms of correct information they contain. To this end, we first show that on a 10-hour ESIC corpus, the ASR errors in the original English speech and its simultaneous interpreting into German and Czech are mutually independent. We then use two sources, English and German, in a multi-source setting for translation into Czech to establish its robustness to ASR errors. Furthermore, we observe this robustness when translating both noisy sources together in a simultaneous translation setting. Our results show that multi-source neural machine translation has the potential to be useful in a real-time simultaneous translation setting, thereby motivating further investigation in this area.

1 Introduction

Speech translation (ST) suffers from automatic speech recognition (ASR) errors, especially in challenging conditions such as non-native language speakers, background noise, named entities and specialized vocabulary usage (Macháček et al., 2019; Gaido et al., 2021, 2022; Anastasopoulos et al., 2022). ASR errors negatively impact translation quality, via the compounding of speech recognition and translation errors (Ruiz et al., 2017; Sperber and Paulik, 2020), thereby limiting the application of automatic speech translation in realistic settings. Fortunately, there are multilingual settings where a source is simultaneously or consecutively interpreted into multiple languages. Many documents are also dubbed or subtitled in offline mode. This simultaneous interpretation takes place either via human interpreters, in the form of dubbing or subtitling. In a situation where the same sentence is available in multiple languages, multi-source machine translation (MT) significantly improves translation quality, especially when the two sources used separately do not yield high quality translations (Dabre et al., 2018; Zoph and Knight, 2016; Nishimura et al., 2018).

Although not yet clearly verified, multi-source MT could be useful in settings where the sources complement each other. In other words, challenges in translation posed by using each source should be independent of one another. Given that ASR is noisy and that multiple sources can help overcome limitations of individual sources, this paper asks the following question: "Can multi-source MT be leveraged for speech translation in a multilingual setting where an original source transcription and its simultaneously interpreted transcription are available?" We decompose this question into three parts where we hypothesize that (a) ASR errors in the original and interpreted transcripts are independent, which makes them complementary, (b) multi-source MT is robust to transcription errors present in individual sources, and (c) the robustness of multi-source MT continues to hold in a simultaneous translation setting. We address each question in Sections 3, 4, and 5, respectively.

To prove our hypotheses, firstly, we verify on the Europarl Simultaneous Interpreting Corpus (ESIC, Macháček et al., 2021) that original speech ASR and interpreted speech ASR are indeed complementary in terms of errors. Secondly, we simulate transcription errors in a full sentence multi-source MT setting for English and German to Czech translation. We clearly show that when both sources are noisy, using them together leads to significant improvements when compared to using them individually, in contrast to drops in quality when at least one of the sources is clean. For example, on ESIC test set with 15% WER noise in English source and 10% WER noise in German source,

Findings of the Association for Computational Linguistics: ACL 2023, pages 3707–3723
July 9-14, 2023

multi-sourcing performs 0.9 BLEU score higher than English source. Finally, we use both sources in a simultaneous translation setting and show that multi-source MT continues to be robust to transcription errors.

Our findings show that multi-source MT has strong potential in a simultaneous translation setting where multiple sources are available via ASR or interpreted ASR. We note that our current analysis is limited to the case where the multiple sources are aligned and hence available at the same time. This is a setting of, e.g., dubbed and subtitled videos where we would want to consider additional target languages. In simultaneous settings where one source is available with a delay, the synchronization of the sources would be a considerable problem, which we leave for future research.

2 Related Work

This paper mainly focuses on ASR errors, multilingual multi-source translation and simultaneous translation.

ASR errors often propagate to MT in cascaded ST systems and in a real time translation setting where ASR systems are used. It is a major issue that affects translation quality. Martucci et al. (2021) propose a method to tune the MT on the training data with artificial noise that mimics ASR errors, via a unigram "lexical noise model" learned on automatic–gold transcript pairs. Other authors propose similar methods for training (Sperber et al., 2017; Di Gangi et al., 2019; Xue et al., 2020; Serai et al., 2022). However, in this work, rather than mimicking ASR noise during training, we complement a noisy source with another whose ASR errors are provably independent of the first. Specifically, we use the lexical noising of Martucci et al. (2021) to simulate noise in multiple source languages and show the robustness of MT models, especially in a multi-source setting.

Multilingualism (Dabre et al., 2020) has been shown to improve translation quality in a variety of situations. In particular, multi-source machine translation has high potential for improving translation quality, but has been relatively underexplored. In the context of multi-source text-to-text translation, Zoph and Knight (2016) and Dabre et al. (2018) showed that leveraging the same sentence in different languages improves translation quality as the two sources are expected to complement hard to translate phenomena in the other source. Although this approach requires multi-parallel sentence-aligned data, the missing sources can be obtained by MT (Nishimura et al., 2018) or via simultaneous interpretation. Rather than training a multi-source model, Firat et al. (2016) propose the "late averaging" which needs multilingual models trained on pairwise bilingual data, which we also focus on when evaluating multi-source models. Late averaging is akin to ensembling via logits averaging, but with source sentences in different languages. These works do not consider transcription errors which are ubiquitous in speech translation, an aspect this paper focuses on.

Multi-sourcing in simultaneous translation settings has not been extensively explored. Dabre et al. (2021) have explored simultaneous multi-pivot translation where a source is translated into a target language via multiple pivot languages, where the pivot languages are translated using multi-source translation. Unlike them, we consider only one pivot language which is interpreted from the source and then use it together with the source to show that the translation quality into the target language improves. Additionally, they do not consider the effect of transcription noise on translation, which we do. Simultaneous translation approaches such as wait-k (Ma et al., 2019) and Local Agreement (LA-n, Polák et al., 2022) are commonly used, and we use the latter for our experiment.

An alternative multi-sourcing approach is to select and use only one source. Macháček et al. (2021) provide an analysis of using either the original, or its simultaneously interpreted equivalent as a source for simultaneous ST. Interpreting is delayed, but shorter and simpler than translationese. Interpreters also segment their speech to sentences differently than the original speakers, so it is not easy to align segments. In any case, selecting sources will involve additional effort and thus we consider using multiple sources together to be a more effective approach. In this regard, multiple language sources, both as text and speech streams, could be used in ASR (Paulik et al., 2005; Soky et al., 2022) as well as in pre-neural MT and ST (Och and Ney, 2001; Paulik and Waibel, 2008; Khadivi and Ney, 2008). Miranda et al. (2013) use them for punctuation restoration. Kocmi et al. (2021) provide a broad analysis of benefits of multilingual MT.

subset	Cs interp.	De interp.	En original
dev	14.84	25.14	13.63
test	14.04	23.79	14.71

Table 1: Transcription WER on ESIC. There are 191 and 179 documents in dev and test subsets. The scores are weighted by number of words in gold transcripts.

3 Parallel Source-Interpreted ASRs are Independent

We assume a multi-source setting with the original speech and its simultaneously interpreted equivalent as the two sources will improve robustness to ASR errors if the errors in the two source streams complement each other. This is not obvious because, on the one hand, the ASRs work independently where they are deployed for different languages, trained on different data, and the processing is fully independent. On the other hand, the content of the speeches is identical. Interpreters' speech pacing also depends on the original speaker, and it may influence the quality of both ASRs the same way. Therefore, in this section, we analyze the dependency of ASR errors in the source and interpreter, on 10-hour ESIC corpus (Macháček et al., 2021) to prove that the ASR errors are indeed independent.

Methodology First, we processed ASR for English original speakers and interpreters into Czech and German. For English, we used the low-latency neural ASR by Nguyen et al. (2021). For German, we used an older hybrid HMM-DNN model trained using the Janus Recognition Toolkit, which features a single-pass decoder (Cho et al., 2013). For Czech, we used Kaldi (Povey et al., 2011) HMM-DNN model trained on Czech Parliament data (Kratochvíl et al., 2020). Table 1 summarizes the transcription quality on ESIC showing that the quality is low, but to the best of our knowledge it is the best one available for this domain.

We then re-used the word alignments of gold transcripts between the original and interpretation as described in Macháček et al. (2021). 38% of tokens were aligned between English and Czech interpretations, and 40% between English and German, see Table 2. It may be caused by the characteristics of the language pair (e.g. compound words in German vs multi-word expressions in English), features of interpreting (non-verbatim translation, shortening) and by errors in automatic alignment. We only analyzed the aligned tokens further. Since there are many tokens left in two 5-hour subsets of the corpus, we consider further analysis as valid.

	En tokens	En-Cs aligned	En-De aligned
dev	44,494	16,962 (38.12%)	17,809 (40.03%)
test	46,151	17,623 (38.19%)	19,280 (41.78%)

Table 2: Number and percentage of aligned tokens in gold transcripts between the original source (English [En]) and its interpretations (German [De] and Czech [Cs]).

En orig.		Cs int. corr.	Cs int. incorr.	De int. corr.	De int. incorr.
dev	**corr.**	13815	1497	7192	1561
	incorr.	1228	422	633	307
test	**corr.**	14204	1655	7895	1638
	incorr.	1344	420	692	336

Table 3: Contingency table of correctly and incorrectly recognized aligned tokens in English source (in rows) and interpretation into Czech and German (in columns), in dev and test subset of ESIC corpus. According to the χ^2 test of statistical independence, in all 4 cases, the parallel recognition is independent with $p < 0.01$.

Finally, we aligned gold and automatic transcripts using Levenshtein edit distance.[1] We classified each token in the ASR transcript as transcribed correctly or not, both for source and interpretations.

Results We made a contingency table (Table 3) and ran a χ^2 test (Pearson, 1900) of statistical independence. The results show that the **parallel source and interpretation ASRs make errors independently** of each other with $p < 0.01$, for both pairs, English-Czech and English-German, for both dev and test subsets.

We manually assessed the severity of the ASR errors and realized that most errors are only in spelling and fluency, and not in adequacy. We therefore conclude that our finding of independence of parallel ASRs may be valid only for ASRs of comparable quality to ours.

4 Multi-Source Speech Translation

Having established that ASR errors are independent, we now analyze whether multi-source neural machine translation (NMT) is robust to noisy sources. We focus on NMT for individual sentences, with gold sentence alignment of the sources and reference. It is a less realistic use-case than translating long speech documents without any sentence segmentation and alignment of the sources, but proving the robustness of multi-sourcing in this

[1] https://pypi.org/project/edlib/

	sent.	doc.	En words	De w.	Cs w.
dev	2002	179	44866	43323	38347
test	1963	189	44273	42491	37695

Table 4: Size statistics of tri-parallel sentence-aligned "revised translations" of ESIC (Macháček et al., 2021). English is original, German and Czech are translations.

setting paves the way for its application in long speech document translation.

Data For training, we use data from OPUS (Tiedemann and Nygaard, 2004), aiming at a multi-way model with English and German on the source side and Czech as a target. We download all the data from OPUS, remove all sentences from IWSLT, WMT, ESIC and other test sets, filter them by language identification, and then process with dual cross-entropy scoring (Junczys-Dowmunt, 2018) using the bilingual NMT models from Tiedemann and Thottingal (2020). We select the top 30 million sentences for each language pair as training data, to prevent overfitting for either. It is also near the threshold that Chen et al. (2021) showed as optimal.

For NMT validation and evaluation, we use the "revised transcript and translations" from ESIC (Macháček et al., 2021). These are the texts that were originally uttered in the European Parliament, transcribed, revised and normalized for reading and publication on the website, and then translated. They are analogous, but not identical, to the gold transcripts of the original and interpretations that we used in Section 3. In addition to the version published by Macháček et al. (2021), we properly align the sentences in all the three languages. Two documents were removed because they missed German translation. The corpus is of comparable size to a usual MT test set. See size statistics in Table 4.

For a contrastive evaluation, we use Newstest11 (Callison-Burch et al., 2011). It contains 3003 sentences in 5 languages: English, German, Czech, French and Spanish, the same amount in each. Newstest11 has references that were translated directly, not through an intermediate language. We also use three additional Czech references of Newstest11 that were translated from German (Bojar et al., 2012).

Multi-Sourcing We convert Marian models to PyTorch to be used with the Hugging Face Transformers (Wolf et al., 2020) library, in which we implement late and early averaging. For both single- and multi-sourcing, we use greedy decoding because beam search support is not implemented with multi-source.

Training details We train a multi-way NMT model using Marian (Junczys-Dowmunt et al., 2018) with English and German as sources, with language identification tokens, and Czech as the target. We use two separate SentencePiece (Kudo and Richardson, 2018) vocabularies, both sizes of 16 000. The source vocabulary is joint for German and English, and the target is only for Czech. The model is a Transformer Base (6 layers, 512 embedding size, 8 self-attention heads, 2048 filter size) trained on 8 Quadro P5000 GPUs with 16 GB memory for 17 days, until convergence.

Checkpoint selection We validate all checkpoints (every 1000 training steps, 15 minutes) on two single sources (English and German) and two multi-sourcing options: early averaging, and late averaging of a single checkpoint with two sources. Furthermore, after the training has ended, we selected top 10 checkpoints that reached the highest BLEU scores for English and German single-source on the ESIC dev set. We evaluated all pairs of the top performing checkpoints in late averaging multi-sourcing setup. The top performing model from all validation and grid search options was selected as a final model. It is late averaging with a pair of distinct checkpoints. We also use these two checkpoints for single source evaluation.

Evaluation Metrics We estimate translation quality by BLEU (Papineni et al., 2002) and chrF2 (Popović, 2016) calculated by sacreBLEU[2] (Post, 2018). We also report the current state-of-the-art metric COMET[3] (Rei et al., 2020) that achieves the highest correlation with direct assessment as a kind of human judgements (Mathur et al., 2020). However, COMET requires one source on the input and is not suitable for multi-source. Therefore, we report it twice (En/De COMET) with two single sources. Note that En COMET scores assume English as source and Czech as target. Since ESIC is tri-parallel, even if the translation is obtained using German or English and German multi-source, we only use the English source as the input to the COMET model. De COMET scores are computed similarly.

[2]Metric signatures: BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp| version:2.2.1, chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1

[3]`wmt20-comet-da` model

Results with clean inputs Table 5 shows the results of multi-sourcing with clean inputs, without any speech recognition noise. One would be tempted to conclude that the translation from English is of a higher quality than the translation from German (e.g. 33 vs. 26 BLEU on ESIC dev set), but such a claim is risky. The metrics measure the match of the candidate translation with the reference sentence (and, in case of COMET, also with the source), and it is conceivable that the English served as the source for the human reference translation. The Czech reference thus may very well exhibit more traits of the English source than of the German source. While the chrF2 scores agree with BLEU, COMET scores seem to indicate that multi-sourcing is as good as, if not better than, using a single source. Since COMET is known to correlate with human judgements better than BLEU (Mathur et al., 2020) our results show that multi-sourcing is indeed a viable solution.

To further shed light on the impact of the source used for creating references, we evaluated the models with Newstest11 and computed the scores with three additional references that were translated only from German. The German single source achieves much higher BLEU than the English source (32.23 vs 16.62 BLEU), with multi-sourcing in between (22.47 BLEU). Similar trends are observed in chrF2 and COMET scores. This is the opposite of ESIC scores, where the reference was obtained from English. It shows that the traits of the source language such as word order, structure of clauses and terms are remarkable in automatic metrics when the reference is constructed from that source, but these effects may be negligible in human evaluation. Appendix C contains more details.

Finally, we consider a "balanced" scenario where an equal number of references comes from both sources and this shows similar scores for both single sources (23.40 vs 22.85 BLEU) with multi-sourcing outperforming them by 0.6 and 1.1 BLEU. We therefore conclude that our multi-source model should be well-prepared for content originating in any of the source languages, but the automatic evaluation metrics may not always capture this. Moving forward, we only use BLEU for simplicity.

Set ref. translation:	**Metric**	**Model** En	De	De+En
ESIC dev En→Cs	BLEU	***33.31**	26.13	*31.90
	chrF2	***60.17**	54.00	*58.59
	En COMET	$^{\times}$**0.920**	0.860	*0.919
	De COMET	$^{\times}$1.007	0.994	***1.022**
ESIC test En→Cs	BLEU	***33.63**	27.99	*32.57
	chrF2	***59.58**	54.75	*58.63
	En COMET	*0.906	0.871	$^{\times}$**0.912**
	De COMET	0.994	$^{\times}$1.006	***1.018**
news11 3×{De→Cs}	BLEU	16.62 ±0.29	**32.23** ±0.53	22.47 ±0.44
	chrF2	44.84 ±0.18	**58.81** ±0.38	49.72 ±0.27
	En COMET	0.528 ±0.002	**0.823** ±0.002	0.652 ±0.003
	De COMET	0.600 ±0.002	**0.967** ±0.001	0.757 ±0.003
news11 {De,En,Fr,Es}→Cs, Cs	BLEU	*23.40	22.85	***23.96**
	chrF2	$^{\times}$**51.00**	50.27	*50.83
	En COMET	0.627	***0.674**	*0.659
	De COMET	0.700	***0.832**	*0.766

Table 5: Evaluation scores with clean inputs (no ASR noise), machine-translated into Czech with single-sourcing English (En) or German (De), or multi-sourcing (De+En), on ESIC and Newstest11 (news11). Newstest is evaluated on a balanced reference that has origin in 5 languages ({De,En,Fr,Es}→Cs translations and Cs original; 600 sentences each), and 3-times with additional references that were translated from German ("3×{De→Cs}"). We report avg±stddev for them. "En COMET" and "De COMET" are run with English and German source, respectively. Maximum scores are in bold. The symbol * means that there is statistically significant difference ($p < 0.05$) from all the lower scores in the same row, $^{\times}$ means no significance (t-test for COMET, paired bootstrap resampling for BLEU and chrF2).

4.1 Modeling Transcription Noise

Although multi-sourcing English and German is not very beneficial when both sources are clean, we hypothesize that it could show benefits with noisy sources. Averaging two noisy sources can lead to cancelling the noise. Since ESIC contains tri-parallel sentence-aligned translations as texts and not speech, and since we want to evaluate different levels of ASR noise, and we do not have many ASRs, we generate the ASR errors artificially.

Custom WER noise model We adopt the lexical noise model by Martucci et al. (2021) and modify it to create outputs with arbitrary WER. The lexical noise model modifies the source by applying insertion, deletion, substitution, or copy operations on each word with probabilities p_I, p_D, and p_S, respectively. The probabilities are learned from the ASR and gold transcript pairs. It thus may learn to shuffle homonyms such as "eight" and "ate".

In the original lexical noise model by Martucci et al. (2021), the target WER is bound to the performance of the given ASR system on which it is trained, and can not be changed. WER is defined as

BLEU	**ESIC dev**	**En WER**								
	single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		$33.3^{\pm0.0}$	$29.7^{\pm0.3}$	$26.3^{\pm0.4}$	$22.9^{\pm0.4}$	$20.4^{\pm0.5}$	$18.2^{\pm0.8}$	$15.8^{\pm0.1}$	$14.0^{\pm0.2}$	$12.1^{\pm0.1}$
De WER 0 %	$26.1^{\pm0.0}$	$31.9^{\pm0.0}$	$30.0^{\pm0.2}$	**$28.5^{\pm0.3}$**	$26.6^{\pm0.1}$	$25.2^{\pm0.4}$	$23.8^{\pm0.3}$	$21.9^{\pm0.3}$	$20.5^{\pm0.2}$	$19.3^{\pm0.3}$
5 %	$23.5^{\pm0.0}$	$30.9^{\pm0.1}$	$29.1^{\pm0.2}$	$27.6^{\pm0.3}$	**$25.7^{\pm0.1}$**	$24.2^{\pm0.4}$	$22.8^{\pm0.4}$	$21.1^{\pm0.4}$	$19.6^{\pm0.2}$	$18.6^{\pm0.2}$
10 %	$21.6^{\pm0.2}$	$30.0^{\pm0.2}$	$28.0^{\pm0.1}$	$26.6^{\pm0.4}$	$24.6^{\pm0.3}$	$23.4^{\pm0.2}$	$21.9^{\pm0.4}$	$20.2^{\pm0.1}$	$18.7^{\pm0.2}$	$17.5^{\pm0.5}$
15 %	$19.0^{\pm0.3}$	$28.9^{\pm0.2}$	$27.1^{\pm0.1}$	$25.7^{\pm0.4}$	$23.7^{\pm0.2}$	$22.4^{\pm0.4}$	$21.0^{\pm0.4}$	$19.3^{\pm0.2}$	$17.8^{\pm0.3}$	$16.7^{\pm0.4}$
20 %	$17.1^{\pm0.3}$	$27.9^{\pm0.4}$	$26.6^{\pm0.2}$	$24.9^{\pm0.4}$	$22.9^{\pm0.1}$	$21.7^{\pm0.5}$	$20.0^{\pm0.4}$	$18.3^{\pm0.2}$	$17.0^{\pm0.1}$	$15.7^{\pm0.1}$
25 %	$15.6^{\pm0.3}$	$27.1^{\pm0.3}$	$25.7^{\pm0.2}$	$24.1^{\pm0.3}$	$22.1^{\pm0.2}$	$20.7^{\pm0.4}$	$19.2^{\pm0.5}$	$17.4^{\pm0.2}$	$16.3^{\pm0.2}$	$14.9^{\pm0.1}$
30 %	$13.8^{\pm0.2}$	$25.9^{\pm0.3}$	$24.5^{\pm0.4}$	$22.8^{\pm0.3}$	$20.9^{\pm0.3}$	$19.6^{\pm0.2}$	$18.3^{\pm0.2}$	$16.3^{\pm0.4}$	$15.1^{\pm0.1}$	$13.9^{\pm0.2}$
35 %	$12.5^{\pm0.2}$	$24.6^{\pm0.4}$	$22.5^{\pm0.4}$	$20.9^{\pm0.2}$	$19.2^{\pm0.1}$	$18.1^{\pm0.5}$	$16.7^{\pm0.3}$	$15.3^{\pm0.3}$	$14.1^{\pm0.2}$	$12.9^{\pm0.1}$
40 %	$10.8^{\pm0.1}$	$23.4^{\pm0.4}$	$21.4^{\pm0.1}$	$20.1^{\pm0.3}$	$18.3^{\pm0.5}$	$17.3^{\pm0.2}$	$16.0^{\pm0.1}$	$14.4^{\pm0.1}$	$13.2^{\pm0.2}$	$12.1^{\pm0.1}$

BLEU	**news11**	**En WER**								
	single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		$23.4^{\pm0.0}$	$21.1^{\pm0.2}$	$19.2^{\pm0.1}$	$17.1^{\pm0.0}$	$15.3^{\pm0.1}$	$13.6^{\pm0.2}$	$12.2^{\pm0.2}$	$10.6^{\pm0.3}$	$9.6^{\pm0.0}$
De WER 0 %	$22.9^{\pm0.0}$	$24.0^{\pm0.0}$	$22.7^{\pm0.0}$	$21.4^{\pm0.2}$	$20.2^{\pm0.1}$	$18.9^{\pm0.2}$	$17.8^{\pm0.1}$	$16.9^{\pm0.1}$	$15.5^{\pm0.1}$	$14.6^{\pm0.2}$
5 %	$20.6^{\pm0.1}$	$23.2^{\pm0.1}$	$21.8^{\pm0.1}$	$20.7^{\pm0.0}$	$19.2^{\pm0.0}$	$18.2^{\pm0.1}$	$17.1^{\pm0.1}$	$16.1^{\pm0.1}$	$14.8^{\pm0.0}$	$13.9^{\pm0.1}$
10 %	$18.8^{\pm0.1}$	$22.5^{\pm0.1}$	$21.3^{\pm0.2}$	$20.1^{\pm0.1}$	$18.6^{\pm0.2}$	$17.7^{\pm0.1}$	$16.4^{\pm0.1}$	$15.5^{\pm0.1}$	$14.1^{\pm0.1}$	$13.2^{\pm0.2}$
15 %	$17.0^{\pm0.3}$	$21.6^{\pm0.2}$	$20.3^{\pm0.2}$	$19.1^{\pm0.2}$	$17.8^{\pm0.1}$	$16.9^{\pm0.1}$	$15.6^{\pm0.0}$	$14.7^{\pm0.1}$	$13.4^{\pm0.1}$	$12.5^{\pm0.1}$
20 %	$15.4^{\pm0.2}$	$20.8^{\pm0.0}$	$19.5^{\pm0.1}$	$18.3^{\pm0.1}$	$17.0^{\pm0.2}$	$16.0^{\pm0.1}$	$14.9^{\pm0.1}$	$14.0^{\pm0.1}$	$12.7^{\pm0.2}$	$12.0^{\pm0.1}$
25 %	$13.8^{\pm0.1}$	$19.9^{\pm0.2}$	$18.7^{\pm0.2}$	$17.7^{\pm0.1}$	$16.3^{\pm0.1}$	$15.4^{\pm0.0}$	$14.0^{\pm0.2}$	$13.2^{\pm0.1}$	$11.9^{\pm0.0}$	$11.1^{\pm0.1}$
30 %	$12.3^{\pm0.3}$	$19.2^{\pm0.3}$	$17.9^{\pm0.2}$	$16.9^{\pm0.3}$	$15.6^{\pm0.1}$	$14.5^{\pm0.2}$	$13.5^{\pm0.3}$	$12.7^{\pm0.2}$	$11.3^{\pm0.1}$	$10.6^{\pm0.1}$
35 %	$11.2^{\pm0.1}$	$18.4^{\pm0.0}$	$17.1^{\pm0.1}$	$16.1^{\pm0.1}$	$15.0^{\pm0.2}$	$13.8^{\pm0.1}$	$12.7^{\pm0.2}$	$11.7^{\pm0.1}$	$10.6^{\pm0.2}$	$9.9^{\pm0.2}$
40 %	$9.9^{\pm0.3}$	$17.1^{\pm0.0}$	$16.1^{\pm0.2}$	$14.9^{\pm0.2}$	$14.0^{\pm0.1}$	$12.9^{\pm0.1}$	$11.7^{\pm0.2}$	$10.7^{\pm0.1}$	$9.9^{\pm0.1}$	$9.1^{\pm0.2}$

Table 6: BLEU (avg±stddev) with transcription noise on ESIC dev set whose reference translations was English and on Newstest11 with balanced reference source language. Green-backgrounded area is where the English single-source outperforms German single-source. Black underlined numbers indicate the area where multi-sourcing achieves higher score than both single-sourcing options. In **bold** is near maximum gap from single-source, more than 2.1 BLEU. Red-colored numbers are where at least one single-source scores higher.

WER	**En**	**De**	**En+De**
15% En, 10% De	23.58±0.16	23.23±0.05	**26.50±0.27**

Table 7: ESIC test multi-sourcing vs single-sourcing BLEU scores on the artificial WER noise level where multi-sourcing achieved the largest improvement.

the number of incorrect words in the ASR transcript divided by the number of correct words in the gold transcript. The errors are either insertions, deletions, or substitutions. In the lexical noise model, insertion is applied independently on the other operations. Therefore, we can decompose WER to the sum of insertion rate and the rate of deletions or substitutions.

In the lexical noise model, the insertion rate equals to the expected number of insertions for each gold word. Since the probability of not inserting is $1 - p_I$, the expected number of repetitions before not inserting succeeds is $\frac{p_I}{1-p_I}$. It is also a mean of a geometric distribution with $p = 1 - p_I$.

The rate of deletions and substitutions is $p_D + (1 - p_D)p_S$, where p_D is the number of deletions. The words that were not deleted can be substituted, and there is $(1 - p_D)p_S$ of them. In summary, the original model WER is

$$\text{WER} = \frac{p_I}{1 - p_I} + p_D + (1 - p_D)p_S, \quad (1)$$

To get a custom target WER, we rescale the learned probabilities by a constant c:

$$\text{WER}_{\text{desired}} = \frac{cp_I}{1 - cp_I} + cp_D + (1 - cp_D)cp_S. \quad (2)$$

We simplify the equation above to

$$\text{WER}_{\text{desired}} \approx cp_I + cp_D + (1 - cp_D)cp_S. \quad (3)$$

It leads to a quadratic function where c can be found easily. Since we work with probabilities, we select the smallest non-negative root as the solution. We release our implementation online.[4]

[4] https://github.com/pe-trik/asr-errors-simulator

Training the noise model For training the noise model, we utilize VoxPopuli (Wang et al., 2021) to retrieve around 100,000 audio and gold transcript sentences in English and 60,000 in German. They are from the same domain as ESIC, both corpora are from the European Parliament. We processed the audio with NVidia NeMo CTC ASRs[5] (Kuchaiev et al., 2019; Gulati et al., 2020). Then we trained the rules of the lexical noise model and applied them on source data. Since the result is deterministic on the random seed of the lexical noise model, we perform multi-sourcing using three different seeds and report average BLEU scores with standard deviation.

Results with transcription noise Table 6 summarizes the BLEU scores of two-source MT with different levels of transcription noise in each of the sources on two sets: ESIC dev with reference translated from English, and Newstest11 with balanced reference. Appendix A contains the corresponding chrF2 scores. Table 7 shows the results on the ESIC test set for the settings where multi-source models achieved the highest improvement due to noisy inputs.

In Table 6, on both sets, we observe that the less noisy single source achieves higher BLEU than the other single source. When the difference in noise levels between the sources is small (close to diagonal in the table), then multi-sourcing reaches slightly higher BLEU than single sources. In case of balanced Newstest11, this area matches the diagonal. In case of ESIC with English original source and reference translated from English, the area of multi-source outperforming single-source is shifted. This tendency is reflected in the test set results in Table 7 as well. Only when the German source is less noisy than the English one, it does improve BLEU in multi-sourcing. We explain it by the discrepancy of source languages for MT and reference that affect BLEU the same way as in offline mode in Section 4. On Newstest11, with the references translated from German, we expect the reverse.

We also observe expected behavior that the more noise, the lower BLEU in all setups. Compare e.g. 33.3 BLEU with zero noise and 12.1 with 40% WER in both sources. With very large noise, it is possible that neither option would be usable. In ESIC dev, e.g. when English WER is 20%, we observe large span, between 5 and 25% WER in German, where multi-sourcing outperforms single source at least by several hundreths of BLEU. This span in Newstest11 is much more narrow, only 20 to 25% WER in German. We hypothesize that it may be caused by the domain difference. The lexical noise model is trained on Europarl. In news domain, there may be fewer words for substitution, so the noise consists more of deletions and insertions, and it might be more harmful for MT in combination of two sources. However, multi-sourcing appears to be robust to ASR errors regardless of whether we have one or both sources as original.

5 Simultaneous Multi-Source

In the previous section, we experimented with offline translation with artificial ASR noise and showed that multi-source models are indeed robust to noise. However, one important use case of speech translation is in a real time setting where simultaneous MT is used. We therefore evaluate the robustness of multi-source models in a simultaneous setting.

5.1 Simultaneous Machine Translation

Simultaneous MT is a task that simulates one subtask of a technology that translates long-form monologue speech in real-time, or with the lowest possible latency. There exist two main approaches to simultaneous MT: streaming and re-translating (Niehues et al., 2018; Arivazhagan et al., 2020). Re-translating systems generate preliminary translation hypotheses that can be updated. Both approaches have complementary benefits and drawbacks. In this paper, we focus on streaming.

We assume that simultaneous MT continuously receives an input text segmented to sentences, one token at a time, as produced by the speaker and upstream tasks. After reading each input token, the system can either produce one or more target tokens, or decide to read the next input token, e.g. to have more context for translation. The goal of simultaneous MT is to translate the input with high quality and low latency. Quality is measured on full sentences as in standard text-to-text MT, e.g. by BLEU. The standard latency measure of simultaneous MT is Average Lagging (AL, Ma et al., 2019). It is an average number of tokens behind an "optimal" policy that generates the target proportionally with reading the source.

Simultaneous MT can be created from standard text-to-text NMT by applying any simultaneous de-

[5] stt_de_quartznet15x5 and stt_en_conformer_ctc_large from https://catalog.ngc.nvidia.com/models

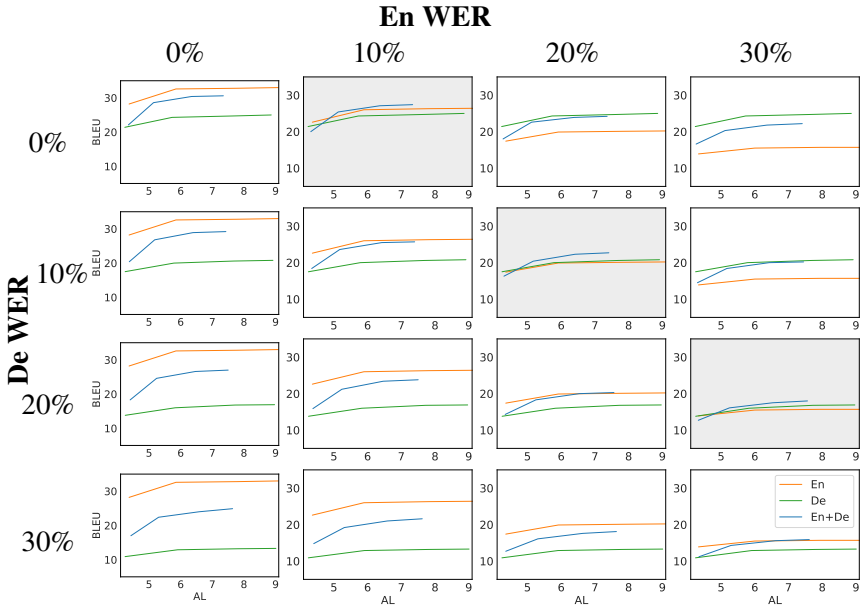

Figure 1: Single-sourcing vs multi-sourcing with different level of artificial ASR noise of the sources (% WER) in simultaneous mode on ESIC dev set. The results are depicted as quality (BLEU) and latency (AL) trade-off of the candidate systems. The plots highlighted by gray background show noise levels where multi-sourcing (En+De, blue line) outperforms both single sources in BLEU at least for AL>5.5.

coding algorithm. However, it is recommendable first to adapt NMT, so it is inclined to translate consecutive sentence prefixes with the same target prefix. We use Local Agreement (LA-n) as a decoding algorithm. It achieved good performance by the best performing system (Polák et al., 2022) in the most recent IWSLT competition (Anastasopoulos et al., 2022). Local Agreement (LA-n) means that n consecutive updates must agree on a target prefix to commit and write. The last committed prefix is then forced as a prefix to decoding the next units. Agreement size n is a parameter that controls the latency.

5.2 Creating Simultaneous MT Systems

In Section 4, we used multi-way models trained on full sentences, but in a simultaneous setting, these models will make mistakes when translating partial sentences using the LA-n approach. Therefore, our multi-way models should first be adapted for partial sentence translation. To this end, we used the multi-way English and German to Czech MT model as a base for simultaneous MT. We fine-tuned the last trained model checkpoint for stable translation on 1:1 mix of incomplete sentence prefixes and full sentences as Niehues et al. (2018). For each source-target pair of the training data, we selected 5-times 1 to 90 % of source and target characters and rounded them to full words. Then, we ran training for 1 day on 1 GPU. We validated BLEU score on ESIC dev and Normalized Erasure (NE, Arivazhagan et al., 2020) on all prefixes of the first 65 sentences (around 1500 words) of ESIC dev set. We ran fine-tuning with multi-way data for English and German as source languages, and for bilingual English-Czech and German-Czech MT.

We stopped training after one day when there were no improvements in stability or quality. Then, we selected one checkpoint for English and one for German that reached acceptable quality and stability values. See Appendix B for details.

5.3 Multi-Sourcing in Simultaneous MT

We use late averaging of the two selected checkpoints for multi-sourcing in simultaneous MT. The only aspects of multi-sourcing in simultaneous mode that differ from single-source or non-simultaneous mode are synchronization of the sources and how to count Average Lagging.

Synchronization In a realistic use-case, it is necessary to synchronize the original speech and simultaneous interpreting. However, we leave it for further work, as our goal is to inspect the limits of multi-sourcing. Therefore, we simulate a case where the sources are optimally synchronized, aligned and parallel to sentence level.

In multi-source mode, we sort all sentence pre-

fixes by proportion of the character length to the sentence length. Each "Read" operation of the multi-source system then receives two prefixes in two languages. One of them is updated by one new token. Every such update is counted to local agreement size. We note that there are other strategies, e.g. count only English source updates to LA-n, but in this paper we have other goal than searching for the best strategy.

AL in multi-source In multi-source setup, we only count Read operations of the English source to AL calculations that we report, and not of the German source because the sources are simultaneous. Counting only German tokens differs negligibly, approximately by 0.1 tokens.

5.4 Simultaneous Multi-Source with Artifical Noise

We want to compare multi-sourcing model to single-sourcing with artificial ASR noise model as in Section 4.1. We evaluate each system on the latency levels with local agreement sizes 2, 5, 10 and 15. Since each evaluation takes approximately 5 hours on 2000 sentences, we report only one run, and not average and deviation on multiple randomly noised inputs.

The results on ESIC dev set are in Figure 1. We can observe the same trends as in the offline case. The single source that is noised less achieves higher BLEU. Multi-sourcing outperforms both single sources when both noise levels are similar and when the English one is lower, e.g. in the case with 10% WER in German and 20% WER in English. We explain it again by the fact that the Czech reference is translated from English, and not German.

Furthermore, on both ESIC and Newstest11 (Figure 1) we observe that multi-sourcing performs worse in the low latency modes, i.e. in AL$<$5 that roughly corresponds to LA$<$5. We assume that the proportional synchronization of the two sources is often inaccurate and may confuse late averaging. In higher latency modes, the synchronization noise at the end of input may be lowered by local agreement. Having validated the multi-source NMT is robust to ASR errors in both full sentence and simultaneous settings, we have paved the way for harder settings where multilingual interpretations of the original source available with different amounts of delay can be used for translation.

6 Conclusion

We have investigated the robustness of multi-source NMT to transcription errors in order to motivate its use in settings where ASRs for the original speech and its simultaneously interpreted equivalent are available. To this end, we first analyzed the 10-hour ESIC corpus and documented that the ASR errors in the two sources are indeed independent, indicating their complementary nature. We then simulated transcription noise for English and German when translating into Czech in single and multi-source NMT settings and observed that using multiple noisy sources is significantly better than individual noisy sources. We then repeated experiments in a simultaneous translation setting and showed that multi-source translation continues to be robust to noise. This robustness of multi-source NMT to noise motivates future research into simultaneous multi-source speech translation, where one source is available with a delay. We will also consider training models with simulated ASR errors to further increase their robustness, especially in multi-source settings.

7 Limitations

Although we have shown the robustness of multi-source NMT to transcription errors in a full-sentence and simultaneous settings, our work has the following limitations:

- Our work does not address the case where the additional source, typically interpreted, is available after a delay. A delayed source may reduce the gains seen by multi-sourcing.
- We have only focused on the Local Agreement (LA-n) approach for simultaneous translation and exploration of other simultaneous approaches such as wait-k remains.
- Human evaluation of translations is pending.
- Evaluation on other language pairs is pending.

Acknowledgements

The research was partially supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation, "Grant Schemes at CU" (reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935), 398120 of the Grant Agency of Charles University, and SVV project number 260 698. Part of the work was done during an internship at NICT.

References

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Ondřej Bojar, Daniel Zeman, Ondřej Dušek, Jana Břečková, Hana Farkačová, Pavel Grošpic, Kristýna Kačenová, Eva Knechtová, Anna Koubová, Jana Lukavská, Petra Nováková, and Jana Petrdlíková. 2012. Additional German-Czech reference translations of the WMT'11 test set. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109, Online. Association for Computational Linguistics.

Eunah Cho, Christian Fügen, Teresa Hermann, Kevin Kilgour, Mohammed Mediani, Christian Mohr, Jan Niehues, Kay Rottmann, Christian Saam, Sebastian Stüker, and Alex Waibel. 2013. A real-world system for simultaneous translation of German lectures. In *Proc. Interspeech 2013*, pages 3473–3477.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Fabien Cromieres, and Sadao Kurohashi. 2018. Exploiting multilingual corpora simply and efficiently in neural machine translation. *Journal of Information Processing*, 26:406–415.

Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. 2021. Simultaneous multi-pivot neural machine translation. *CoRR*, abs/2104.07410.

Matti Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019. Robust neural machine translation for clean and noisy speech transcripts. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Who are we talking about? handling person names in speech translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 62–73, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Marco Gaido, Susana Rodríguez, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2021. Is "moby dick" a whale or a bird? named entities and terminology in speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1707–1716, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast

neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Shahram Khadivi and Hermann Ney. 2008. Integration of speech recognition and machine translation in computer-assisted translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1551–1564.

Tom Kocmi, Dominik Macháček, and Ondřej Bojar. 2021. *The Reality of Multi-Lingual Machine Translation*, volume 22 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Prague, Czechia.

Jonáš Kratochvíl, Peter Polák, and Ondřej Bojar. 2020. Large corpus of Czech parliament plenary hearings. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6363–6367, Marseille, France. European Language Resources Association.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Dominik Macháček, Jonáš Kratochvíl, Tereza Vojtěchová, and Ondřej Bojar. 2019. A speech test set of practice business presentations with additional relevant texts. In *Statistical Language and Speech Processing*, pages 151–161, Cham. Springer International Publishing.

Dominik Macháček, Matúš Žilinec, and Ondřej Bojar. 2021. Lost in Interpreting: Speech Translation from Source or Interpreter? In *Proc. Interspeech 2021*, pages 2376–2380.

Giuseppe Martucci, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. Lexical Modeling of ASR Errors for Robust Speech Translation. In *Proc. Interspeech 2021*, pages 2282–2286.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

João Miranda, João Paulo Neto, and Alan W Black. 2013. Improved punctuation recovery through combination of multiple speech streams. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 132–137.

Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. 2021. Super-Human Performance in Online Low-Latency Recognition of Conversational Speech. In *Proc. Interspeech 2021*, pages 1762–1766.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-Latency Neural Speech Translation. In *Proc. Interspeech 2018*, pages 1293–1297.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with data augmentation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 48–53, Brussels. International Conference on Spoken Language Translation.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

M. Paulik, S. Stuker, C. Fugen, T. Schultz, T. Schaaf, and A. Waibel. 2005. Speech translation enhanced automatic speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 121–126.

Matthias Paulik and Alex Waibel. 2008. Extracting clues from human interpreter speech for spoken language translation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5097–5100.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. CUNI-KIT system

for simultaneous speech translation task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, K. Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. In *Proceedings of ASRU 2011*, pages 1–4. IEEE Signal Processing Society.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Nicholas Ruiz, Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico. 2017. Assessing the Tolerance of Neural Machine Translation Systems Against Speech Recognition Errors. In *Proc. Interspeech 2017*, pages 2635–2639.

Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. 2022. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:890–900.

Kak Soky, Sheng Li, Masato Mimura, Chenhui Chu, and Tatsuya Kawahara. 2022. Leveraging simultaneous translation for enhancing transcription of low-resource language via cross attention mechanism. In *Interspeech*.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96, Tokyo, Japan. International Workshop on Spoken Language Translation.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haiyang Xue, Yang Feng, Shuhao Gu, and Wei Chen. 2020. Robust neural machine translation with ASR errors. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 15–23, Seattle, Washington. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

A chrF2 Scores with Noisy Inputs

There is an evidence that chrF2 correlates with human judgements better than BLEU. In Table 8, we can see that for multi-sourcing with noisy inputs on ESIC dev, chrF2 are indeed higher than single-sourcing, and this correlates with the BLEU score gains in Table 6. On the other hand, for Newstest11, chrF2 scores do not indicate any improvements. While the corresponding BLEU scores in

Table 6 indicated improvements of multi-sourcing with noisy inputs, the magnitude of these gains were minor, much smaller than those observed for ESIC. This gives us sufficient reason to believe that multi-sourcing should be useful in a setting like ESIC, where the reference is created from only one source, which is more realistic than the "balanced" use-case of Newstest11, where the reference originates from 5 languages.

B Checkpoint Selection for Simultaneous Multi-Source

The checkpoints that we selected for simultaneous multi-source decoding (recall Section 5.2) was the multi-way checkpoint for English and bilingual one for German. Table 9 summarizes the results of fine-tuning for stability. BLEU decreased marginally (by 0.2 on English and 0.9 on German), while normalized erasure (NE) dropped by 40% in English and 52% on German.

Based on some outputs, we explain higher NE in German-to-Czech by discrepancy in word orders. Many erasures were caused by an incorrect presumption of the final verb. Regardless, our fine-tuned models exhibit significantly reduced NE and can be reliably used for simultaneous translation using the LA-n approach.

C Effect of Reference Source Language

To explain the effect of reference source language, we run a contrastive evaluation on the subset of Newstest11 that consists only from the documents that originate in English. We compare BLEU measures with a reference translated directly from Czech, and with three additional references translated only from German (Bojar et al., 2012).

The results of simultaneous mode (recall Section 5) are in Figure 2. We observe the same trends as in offline mode in Section 4. The BLEU score is higher for the single source with the language from which the reference was translated. When this source is noised substantially more than the other, multi-sourcing outperforms both by a small margin.

In case of German references, the nearest margin to single-sourcing is much smaller than with the English references. We assume it is because the structural difference of English source and German-Czech references is larger than German to English-Czech reference. It is documented also by BLEU scores with zero noise (33 and 20 on references from English vs 16 and 30 on references from German).

chrF2	**ESIC dev**	**En WER**								
	single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		$60.2^{\pm0.0}$	$57.2^{\pm0.2}$	$54.4^{\pm0.2}$	$51.4^{\pm0.3}$	$49.2^{\pm0.7}$	$46.8^{\pm0.8}$	$44.3^{\pm0.0}$	$42.1^{\pm0.3}$	$40.1^{\pm0.0}$
0 %	$54.0^{\pm0.0}$	$58.6^{\pm0.0}$	$56.9^{\pm0.2}$	$55.6^{\pm0.1}$	$53.7^{\pm0.2}$	$52.3^{\pm0.5}$	$50.9^{\pm0.4}$	$49.2^{\pm0.3}$	$47.5^{\pm0.2}$	$46.1^{\pm0.2}$
5 %	$51.8^{\pm0.1}$	$57.7^{\pm0.1}$	$56.2^{\pm0.2}$	$54.8^{\pm0.1}$	$52.9^{\pm0.2}$	$51.4^{\pm0.6}$	$50.0^{\pm0.4}$	$48.3^{\pm0.3}$	$46.7^{\pm0.3}$	$45.4^{\pm0.2}$
10 %	$49.9^{\pm0.2}$	$56.8^{\pm0.2}$	$55.1^{\pm0.1}$	$53.7^{\pm0.3}$	$51.8^{\pm0.3}$	$50.4^{\pm0.3}$	$49.0^{\pm0.4}$	$47.3^{\pm0.1}$	$45.6^{\pm0.2}$	$44.3^{\pm0.2}$
15 %	$47.6^{\pm0.3}$	$55.8^{\pm0.0}$	$54.2^{\pm0.1}$	$52.8^{\pm0.3}$	$50.9^{\pm0.2}$	$49.6^{\pm0.4}$	$48.1^{\pm0.5}$	$46.4^{\pm0.3}$	$44.9^{\pm0.3}$	$43.6^{\pm0.1}$
20 %	$45.7^{\pm0.3}$	$54.9^{\pm0.2}$	$53.5^{\pm0.1}$	$51.9^{\pm0.3}$	$50.2^{\pm0.1}$	$48.7^{\pm0.6}$	$47.2^{\pm0.4}$	$45.4^{\pm0.2}$	$43.9^{\pm0.3}$	$42.6^{\pm0.3}$
25 %	$44.0^{\pm0.4}$	$54.2^{\pm0.4}$	$52.9^{\pm0.1}$	$51.3^{\pm0.2}$	$49.3^{\pm0.2}$	$48.1^{\pm0.4}$	$46.5^{\pm0.4}$	$44.7^{\pm0.2}$	$43.3^{\pm0.2}$	$41.7^{\pm0.0}$
30 %	$42.1^{\pm0.3}$	$53.1^{\pm0.3}$	$51.7^{\pm0.3}$	$50.2^{\pm0.2}$	$48.3^{\pm0.3}$	$46.8^{\pm0.3}$	$45.4^{\pm0.5}$	$43.5^{\pm0.3}$	$42.2^{\pm0.2}$	$40.6^{\pm0.1}$
35 %	$40.5^{\pm0.2}$	$52.0^{\pm0.3}$	$50.0^{\pm0.3}$	$48.7^{\pm0.2}$	$46.9^{\pm0.1}$	$45.7^{\pm0.5}$	$44.1^{\pm0.4}$	$42.4^{\pm0.2}$	$41.0^{\pm0.1}$	$39.7^{\pm0.1}$
40 %	$38.6^{\pm0.2}$	$51.1^{\pm0.2}$	$49.2^{\pm0.2}$	$47.8^{\pm0.3}$	$46.0^{\pm0.4}$	$44.8^{\pm0.3}$	$43.2^{\pm0.4}$	$41.5^{\pm0.1}$	$39.8^{\pm0.3}$	$38.6^{\pm0.1}$

De WER

chrF2	**news11**	**En WER**								
	single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		$51.0^{\pm0.0}$	$48.8^{\pm0.2}$	$46.9^{\pm0.1}$	$44.9^{\pm0.1}$	$43.0^{\pm0.1}$	$41.0^{\pm0.1}$	$39.4^{\pm0.1}$	$37.3^{\pm0.1}$	$35.8^{\pm0.0}$
0 %	$50.3^{\pm0.0}$	$50.8^{\pm0.0}$	$49.5^{\pm0.0}$	$48.0^{\pm0.1}$	$46.7^{\pm0.1}$	$45.3^{\pm0.2}$	$43.8^{\pm0.3}$	$42.6^{\pm0.2}$	$41.0^{\pm0.1}$	$39.7^{\pm0.1}$
5 %	$48.3^{\pm0.1}$	$50.0^{\pm0.0}$	$48.6^{\pm0.0}$	$47.3^{\pm0.0}$	$45.7^{\pm0.1}$	$44.5^{\pm0.1}$	$43.1^{\pm0.1}$	$41.8^{\pm0.0}$	$40.1^{\pm0.1}$	$38.8^{\pm0.1}$
10 %	$46.5^{\pm0.2}$	$49.2^{\pm0.1}$	$47.9^{\pm0.2}$	$46.4^{\pm0.1}$	$44.8^{\pm0.1}$	$43.8^{\pm0.0}$	$42.3^{\pm0.0}$	$40.9^{\pm0.1}$	$39.2^{\pm0.2}$	$38.0^{\pm0.1}$
15 %	$44.7^{\pm0.2}$	$48.1^{\pm0.1}$	$46.7^{\pm0.1}$	$45.4^{\pm0.1}$	$43.9^{\pm0.1}$	$42.8^{\pm0.0}$	$41.2^{\pm0.0}$	$40.1^{\pm0.1}$	$38.4^{\pm0.0}$	$37.1^{\pm0.0}$
20 %	$42.9^{\pm0.1}$	$47.1^{\pm0.0}$	$45.8^{\pm0.1}$	$44.3^{\pm0.0}$	$42.9^{\pm0.1}$	$41.7^{\pm0.1}$	$40.4^{\pm0.1}$	$39.1^{\pm0.0}$	$37.4^{\pm0.1}$	$36.3^{\pm0.1}$
25 %	$41.1^{\pm0.1}$	$46.1^{\pm0.2}$	$44.8^{\pm0.0}$	$43.6^{\pm0.1}$	$42.0^{\pm0.1}$	$40.8^{\pm0.1}$	$39.3^{\pm0.2}$	$38.1^{\pm0.1}$	$36.4^{\pm0.1}$	$35.3^{\pm0.1}$
30 %	$39.4^{\pm0.2}$	$45.3^{\pm0.3}$	$43.9^{\pm0.2}$	$42.6^{\pm0.2}$	$41.1^{\pm0.1}$	$39.9^{\pm0.2}$	$38.5^{\pm0.2}$	$37.3^{\pm0.1}$	$35.7^{\pm0.0}$	$34.5^{\pm0.2}$
35 %	$38.0^{\pm0.2}$	$44.3^{\pm0.2}$	$42.9^{\pm0.3}$	$41.5^{\pm0.1}$	$40.2^{\pm0.2}$	$38.9^{\pm0.2}$	$37.6^{\pm0.1}$	$36.4^{\pm0.1}$	$34.9^{\pm0.0}$	$33.7^{\pm0.2}$
40 %	$36.2^{\pm0.2}$	$43.2^{\pm0.2}$	$41.9^{\pm0.2}$	$40.5^{\pm0.2}$	$39.1^{\pm0.1}$	$37.9^{\pm0.1}$	$36.4^{\pm0.2}$	$35.2^{\pm0.1}$	$33.8^{\pm0.2}$	$32.8^{\pm0.2}$

De WER

Table 8: chrF2 (avg±stddev) with transcription noise on ESIC dev set whose reference translations was English and on Newstest11 (news11) with balanced reference source language. The area with the green background is where the English single-source outperforms German single-source. Black underlined numbers indicate the area where multi-sourcing achieves higher score than both single-sourcing options. Red-colored numbers are where at least one single-source scores higher.

	En		De	
checkpoint	BLEU	NE	BLEU	NE
starting	33.2	1.77	25.9	3.15
selected	33.0	1.21	25.0	1.52
diff	-0.2	-40%	-0.9	-52%

Table 9: The results of fine-tuning for stability.

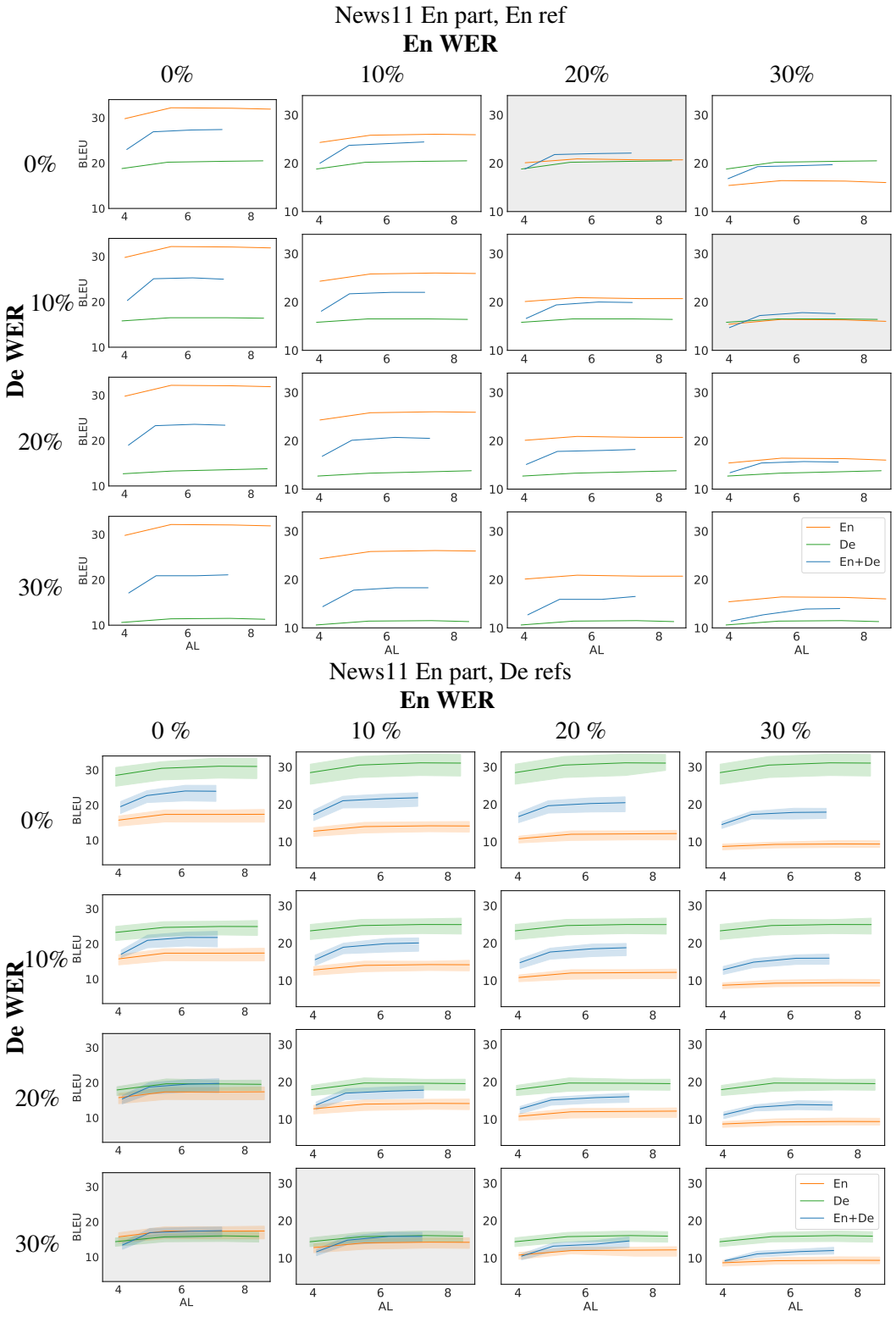

Figure 2: Single-sourcing vs multi-sourcing with different level of artificial ASR noise of the sources (% WER) in simultaneous mode on Newstest11 subset (598 sentences) originally in English. In the upper grid, the Czech reference is translated from English, while in the lower, there is average and standard deviation of BLEU counted against the 3 additional references translated from German (Bojar et al., 2012). Grey highlighting indicates area where multi-sourcing (En+De, blue line) outperforms or is on-par with both single sources in BLEU.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2-5

☑ B1. Did you cite the creators of artifacts you used?
2-5

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
we didn't create anything from potentially harmful data. The cited authors are responsible for the data we used.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2-5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

2-5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
2-5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2-5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
2-5

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.