It is Not as Good as You Think!
Evaluating Simultaneous Machine Translation on Interpretation Data

Jinming Zhao[1] **Philip Arthur**[2] **Gholamreza Haffari**[1] **Trevor Cohn**[3] **Ehsan Shareghi**[1,4]

[1]Department of Data Science & AI, Monash University

[2]Oracle Digital Assistant, Oracle Corp. [4]Language Technology Lab, University of Cambridge

[3]School of Computing and Information Systems, The University of Melbourne

first.last@{monash.edu, unimelb.edu.au, oracle.com}

Abstract

Most existing simultaneous machine translation (SiMT) systems are trained and evaluated on offline translation corpora. We argue that SiMT systems should be trained and tested on real interpretation data. To illustrate this argument, we propose an interpretation test set and conduct a realistic evaluation of SiMT trained on offline translations. Our results, on our test set along with 3 existing smaller scale language pairs, highlight the difference of up-to 13.83 BLEU score when SiMT models are evaluated on translation vs interpretation data. In the absence of interpretation training data, we propose a translation-to-interpretation (T2I) style transfer method which allows converting existing offline translations into interpretation-style data, leading to up-to 2.8 BLEU improvement. However, the evaluation gap remains notable, calling for constructing large-scale interpretation corpora better suited for evaluating and developing SiMT systems. [1]

1 Introduction

Simultaneous interpretation (SI) is a task of translating natural language in real time. SiMT systems are expected to generate interpreted text as if the text was produced by human interpreters while maintaining acceptable delay (Ma et al., 2019; Arthur et al., 2021). However, most current SiMT systems are trained and evaluated on offline translations differing from real-life SI scenarios where translations are flexibly paraphrased, without compromising the source message (He et al., 2016; Paulik and Waibel, 2009). For instance, in Table 1 the interpretation sentence drops "at this point" and condenses "seriousness of this line of argument" to "agreement"; it delivers the source message as reliably as the offline translation.

Source: Ich werde (*I'm*) an diesem Punkt (*at this point*) darauf verzichten, (*refrain from*) einen Kommentar zur Ernsthaftigkeit dieser Argumentationsweise (*to comment on the seriousness of this line of reasoning*) abzugeben.
Offline Translation: (At this point,) I will refrain from commenting on the seriousness of this line of argument.
Interpretation: I'm not going to comment on that agreement.

Table 1: Translation and interpretation differ in style while conveying the same source information.

Prior work attempted to build interpretation corpora in a small scale (Tohyama and Inagaki, 2004; Shimizu et al., 2014; Bernardini et al., 2016), or constructed speech interpretation training corpora for MT tasks (Paulik and Waibel, 2010). But, very little attempt has been made on empirically quantifying the evaluation gap. An exception is Shimizu et al. (2013) which incorporated interpretation data in the training stage of a statistical MT system, but the lack of training data and the scale of evaluation set resulted in a marginal BLEU score difference.[2]

We compile a genuine interpretation test set of 1k utterances from the European Parliament (EP) Plenary focusing on German→English. We examine the real performance gap of wait-k (Ma et al., 2019), a state-of-the-art SiMT system, on our test set along with 3 smaller scale (Bernardini et al., 2016) translation and interpretation language-pairs and observe a drop of up-to 13.83 BLEU score. In the absence of interpretation-style training data, we propose a simple and effective translation-to-interpretation (T2I) style transfer method to produce pseudo-interpretations from abundant offline translations. Training on our T2I transferred data, we observe an improvement of ∼2.8 BLEU score. Our findings necessitate further developments towards constructing large-scale interpretation corpora, designing domain adaptive techniques and models more reflective of real-life interpretations.

[1]Our annotated test sets are available at https://github.com/mingzi151/InterpretationData.

[2]Concurrently, Zhang et al. (2021) trained a system on an offline corpus and evaluated on interpretation test sets, not available to the public at the time of writing our paper.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6707–6715
November 7–11, 2021.

2 German→English Interpretation Data

We provide an overview of our data construction and move full details in *Appendix A.1*.

Collection. We crawled data from the EP Plenary[3] between 2008 and 2012[4] and downloaded 238 debates consisting of speech transcriptions, offline translations and interpretation videos. We used Google speech API to transcribe the interpretation videos and normalize automatic speech recognition (ASR) outputs, yielding 323-hour of transcriptions.

Cleaning, Alignment, and Segmentation. We removed duplicates and the dialogues with non-German source sentences, while using available offline translations to retrieve named entities; this resulted in 5,239 dialogues. We filtered out dialogues with interpretations less than 4 words, and call the resulting interpretations Raw hereafter.

We further removed cases whose sources contained either (1) less than 20 tokens, (2) less than 150 words and included pre-defined signals, or (3) a different number of sentences from the corresponding offline translations. and whose sources and offline translations had a different number of sentences. Next a manual process was applied, including removals of dialogues with non-essential contents and truncation of interpretations whose first and last sentences did not match the corresponding offline translations (mostly due to imperfect audio segmentation). 987 dialogues[5] were thus retained, each of which having 14.5 sentences on average.

We aligned translations with transcriptions (interpretations). For each dialogue, as the transcriptions may not be well segmented in the ASR process, we identified sentences in the transcriptions with stanza (Qi et al., 2020), before segmenting them using dynamic programming. Manual inspection revealed that there were a portion of mismatched pairs, which was due to occasional interpreting failure resulting from interpreters' accumulated cognitive load (Mizuno, 2017; Sudoh et al., 2020). We further removed pairs the lengths of whose source and target were far off, and call it Clean, containing triples <source, translation, interpretation>.

[3] www.europarl.europa.eu/plenary/en/debates-video.html

[4] Beyond this period, offline translations are not provided.

[5] One dialogue is attached in *Appendix A.1*. Note each source sentence and offline translation in the dialogues may consist of several sentences.

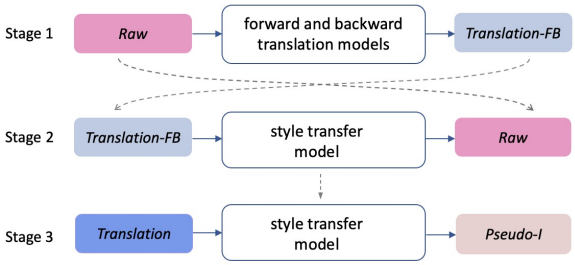

Figure 1: T2I style transfer in unsupervised settings.

Translation and Interpretation Test Sets. To ensure the quality of interpretation data for evaluation, we hired a bilingual German-English speaker to annotate a randomly selected subset (107 dialogues) of the 987 dialogues in two stages: segmentation and ASR error correction. This gave us two versions of test set: Interpretation^{ASR}, Interpretation.

In the first stage, the annotator was asked to match the correct target sentence(s) against each source sentence. The annotator was asked to find interpretation text for each German sentence, when impossible, multiple sentences were allowed. Additionally, to comply with human speaking styles, we allowed minor omissions of unimportant English texts as long as the main idea of German text was conveyed (such as conjunctions). In the second stage, the annotator was instructed to correct ASR errors while applying minimal changes to the sentences.

Ultimately, our test sets comprise 1,090 triples of <source, translation, interpretation> which were further cross checked to enforce quality control.

3 T2I Style Transfer

Offline translated texts and online interpreted texts differ in various aspects, including lengths, sentence structure and lexicon; this is fundamentally contributed by the fact that interpreters use tactics to minimize delay and reduce the load of retention (Mizuno, 2017; Camayd-Freixas, 2011). For example, interpreters tend to break a source sentence into several smaller chunks (see He et al. (2016) for more tactics). Yet, while exhibiting stylistic differences, both preserve the key source message. As will be seen (§4.3) these differences amount to a significant evaluation gap.

While the ideal solution is using human annotators to create interpretation training corpora, in the absence of resources, we propose a simple technique, T2I style transfer, to convert existing

translation data into interpretation-style data with a style transfer model. Training such a model would require paired translations and interpretations, which are not available in large quantities. Rather, our proposed approach allows fulfilling the goal of style transferring abundant translation data to interpretation-like data both in supervised settings, where Clean is leveraged, and unsupervised settings, where only Raw is used.

Supervised training Given that our Clean set consists of roughly 4.2k triples, we opt for statistical MT systems which inherently require far less data for sequence-to-sequence mapping tasks compared to their neural counterparts. Furthermore, conducting style transfer in the same language involves word replacement and ordering, which conforms with the behaviors of SMT systems that chunk an input sequence into segments, translate, and reorder the translated chunks (Lopez, 2008). More specifically, we employ two classic statistical MT methods: phrase-based SMT (PBMT) (Koehn et al., 2003) and Hierarchical phrase-based MT (HPBMT) (Chiang, 2005).[6] A similar framework was tried by Xu et al. (2012) for text simplification.

We will describe the T2I pipeline process for unsupervised settings, as both settings have a similar process with different data configurations. The main difference is that we use Clean instead of Raw, which will be detailed in §4.1.

Unsupervised training Figure 1 shows the three stages of our T2I approach in unsupervised settings: the first stage is to convert interpretations in Raw to translation-style data by applying round-trip translation on interpretations, with pretrained NMT models (Ng et al., 2019). It is expected that the outputs after this round-tripping, denoted as Translation-FB, sit close to the translation domain, thus achieving the effects of interpretation-to-translation. The second stage is to train a style transfer model to learn the mapping between the data points in Translation-FB and their corresponding interpretations in Raw. Lastly, we apply the trained style transfer model on offline Europarl translations and produce interpretation-like sequences which we call Pseudo-I.[7]

[6]PBMT creates a phrase table, a reordering model and a language model, followed by tuning their weights with MERT on parallel data. HPBMT leverages both phrase-based translation and syntax-based translation, and operates on context-free grammar rules.

[7]Examples of Pseudo-I are provided in *Appendix A.2*.

4 Experiments

In this section, we present datasets details (§4.1) followed by the descriptions of our baselines and style transfer models (§4.2). We report results by underlining the performance gap between evaluation on translated and interpreted texts (§4.3), and showing the effectiveness of our T2I style transfer both quantitatively and qualitatively (§4.4).

We followed the instructions in Arthur et al. (2021) to preprocess data, and their hyperparameters for training all wait-k models. For style transfer models, we used the standard setup for both PBMT and HPBMT.[8]

4.1 Datasets

We conducted evaluation investigation on four languages pairs, including German (DE), French (FR), Polish (PL), Italian (IT) → English (EN) , and used Europarl v7 corpus (Koehn, 2005) for training a SiMT model for each pair (see Table 2 for data statistics). For DE-EN, our annotated test set has 1,051 triples, for Interpretation^{ASR} and Interpretation. For the rest, we used EPTIC (Bernardini et al., 2016), a small-scale parallel corpus with data collected from the EP Plenary; it has source languages of FR, PL and IT, with 675, 463 and 480 instances, respectively.

In the experiments of bridging the evaluation gap, Raw has 120,114 and 1,000 utterances for training and dev sets, while Clean has 4,240 triples, all used for training style transfer models. To train PBMT, we augmented Clean by forward translating its source-side data to the target language, together with EPTIC, while using EPTIC to select the best weights for PBMT. We deployed the trained style transfer models on translations of Europarl (DE-EN) to get Pseudo-I. Pairing it with source sentences of Europarl gives us style transferred Europarl.

4.2 Model

Baseline We used wait-k (with k=3) as SiMT systems for its simplicity and effectiveness (Ma et al., 2019). We compared the following wait-k baselines: i) trained on Europarl; ii) adapted on Raw. Performance was evaluated by BLEU[9], average proportion (AP) and lagging (AL) (Cho and Esipova, 2016; Ma et al., 2019). AP measures the

[8]http://www.statmt.org/moses/?n=Moses.Overview

[9]https://github.com/mjpost/sacreBLEU

	# of pairs			Evaluation					
	Europarl Offline		⋆	Translation Test			Interpretation Test		
Lang.	Train	Dev	Test	AP	AL	Bleu	AP	AL	Bleu
DE	1,666,904	3,587	1,051^{+}	0.61	2.84	22.78	0.61	2.84	12.34
FR	1,929,486	9,736	675	0.58	2.41	21.24	0.58	2.41	9.28
PL	601,021	2,035	463	0.61	2.94	24.24	0.61	2.94	13.71
IT	1,832,809	9,256	480	0.56	2.45	24.47	0.56	2.45	10.64

Table 2: Data statistics (# pairs) and evaluation gap using translation vs interpretation test set. (⋆) Test data for both translation and interpretation sets for FR, PL, and IT were from EPTIC, and DE was our proposed set. (+) We further removed 39 cases from our 1090 triples which heavily overlapped with the training data. All training data were Europarl offline translations.

			BLEU		
Model	AL	AP	Translation	Interpretation^{ASR}	Interpretation
Europarl					
train on <Source, Translation>	0.61	2.84	22.78*	11.47*	12.34*
adapt on <Source-FB, Raw>	<u>0.61</u>	<u>2.76</u>	20.71	12.05	12.69
Style transferred Europarl					
train on <Source, Pseudo-I>					
Seq2Seq (unsupervised)	0.66	4.45	10.42	7.79	10.33
HPBMT (unsupervised)	0.61	2.92	18.80	13.53	13.21
PBMT (supervised)	0.61	2.93	17.34	13.87	13.56
HPBMT (supervised)	0.62	3.00	18.55	**14.26**	**13.60**

Table 3: Evaluation on human annotated Translation Test, Interpretation Test^{ASR} and Interpretation Test. *: performance gap. <u>Underlined</u>: lowest delay across systems. **Bold**: Best BLEU on Interpretation Test.

percentage of read source tokens for every generated target token, while AL measures the number of lagged source tokens until all source tokens are read.

Style Transfer Models In supervised settings, we used PBMT and HPBMT; in unsupervised settings we only used HPBMT, as PBMT requires additional paired data to find the best weights. We deployed Moses (Koehn et al., 2007) for above systems. We also experimented with a Seq2Seq (unsupervised) model (Ott et al., 2019) to compare.

4.3 Performance Gap

We train separate wait-k models for the four language pairs and report the evaluation results on their corresponding Translation Test and Interpretation Test[10] in Table 2. The observed significant gap of up-to 13.83 BLEU score (24.47 vs 10.64 for IT) highlights the daunting task SiMT models face in real-life SI. Interestingly, the gap for DE-EN is the lowest, and this is likely to be due to the fact that both are Germanic languages.

We explored the feasibility of narrowing the performance gap using our T2I method on DE-EN. Being a head-final language, German is more difficult to interpret than head-initial languages (e.g., EN, FR, IT and PL), and interpreters must hold information until verb phrases are heard (Mizuno, 2017). Furthermore, having created the datasets for German, our experimental setup was year/domain-consistent for training the baselines and style-transfer models, which allows us to isolate if the improvement was purely achieved by our T2I transfer method.

Full results are reported in Table 3. When wait-k was adapted on Source-FB, Raw, the lowest delay was seen, implying using interpretation corpora is effective in reducing delay. Translation quality can be further boosted with our style transfer method, as discussed next.

[10] A one-to-one correspondence exists between both sets for all language pairs.

Source	Gold	wait-k	wait-k + T2I
Es erfüllt mich mit großer Traurigkeit.	It is with great sadness.	I am with great sadness.	I'm very sorry about that.
Der Bericht begrüßt außerdem ausdrücklich den Vorschlag der Kommission für eine horizontale Richtlinie zum Thema Antidiskriminierung.	The report also explicitly welcomes the Commission's proposal for a horizontal directive covering all forms of discrimination.	The report also expressly welcomes the Commission's proposal for a horizontal directive on the subject of anti-discrimination.	The report welcome the commission's proposal for a horizontal directive on anti-discrimination legislation.

Table 4: Examples of translation predicted by wait-k and translation predicted by a style transferred model, along with their source sentences and gold-translation.

4.4 Impacts of T2I Style Transfer

Quantitative Analysis Our approach yields significantly better results on Interpretation^{ASR} compared to baselines. Our best model outperformed pre-trained wait-k by 2.79 BLEU score. On Interpretation, we see a similar trend but with a smaller margin. We speculate the drop occurred because the T2I models were trained on ASR outputs, which is in the same domain as targets of Interpretation^{ASR}.

Nevertheless, all T2I models work consistently well in supervised and unsupervised settings. Moreover, our approach surpasses Seq2Seq by 6.47 points on Interpretation^{ASR}, verifying that in low-resource settings SMT is superior to NN. Our results, including adapting wait-k on Raw and using T2I to create training corpus, suggest that adequate numbers of paired translation, clean interpretation would lead to decreased delay and better translation quality.

The limitation, however, is that the BLEU score still remains relatively low, which is not surprising, for we only used a minimal number of parallel data in the style transfer process. Hence, while our method does not remove the performance gap, it can still serve as a data augmentation technique to complement future interpretation training data.

Qualitative Analysis To compare translations produced by the vanilla wait-k and its variants trained on T2I transferred data, we give examples in Table 4 along with their sources and gold translations. In the first example, T2I variant is colloquial, implying interpreters giving up the original words and restating the source message (Camayd-Freixas, 2011). T2I variant in the other example is a more condensed translation by dropping unimportant words (Sudoh et al., 2020), such as "also expressly" and "the subject of". Both examples confirm human interpreters' tactics (He et al., 2016).

5 Conclusion

We investigated the SiMT evaluation gap when SiMT models were tested on interpretation vs translation, across four language pairs. To the best of our knowledge, this is the first work quantifying this gap empirically. To bridge the gap, we proposed a data augmentation style transfer technique to create parallel pseudo-interpretations from abundant offline translation data. Our results show an improvement of 2.8 BLEU score. We hope our work and the highlighted evaluation discrepancy can encourage further developments of datasets and models more reflective of real-world SI scenarios.

6 Acknowledgements

This work is supported by the ARC Future Fellowship FT190100039 and an Amazon Research Award to G.H. We would like to thank anonymous reviewers for their valuable comments.

References

Philip Arthur, Trevor Cohn, and Gholamreza Haffari. 2021. Learning coupled policies for simultaneous machine translation using imitation learning. In *Proceedings of EACL*.

Silvia Bernardini, Adriano Ferraresi, and Maja Miličević. 2016. From epic to eptic—exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1):61–86.

Erik Camayd-Freixas. 2011. Cognitive theory of simultaneous interpreting and training. In *Proceedings of AMTA*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *CoRR*, abs/1606.02012.

He He, Jordan L. Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of NAACL HLT*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

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 ACL*.

Akira Mizuno. 2017. Simultaneous interpreting and cognitive constraints. *Bull. Coll. Lit*, 58:1–28.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's WMT19 news translation task submission. In *Proceedings of WMT*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT*.

Matthias Paulik and Alex Waibel. 2009. Automatic translation from parallel speech: Simultaneous interpretation as MT training data. In *IEEE Workshop on Automatic Speech Recognition & Understanding*.

Matthias Paulik and Alex Waibel. 2010. Spoken language translation from parallel speech audio: Simultaneous interpretation as SLT training data. In *Proceedings of ICASSP*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the ACL: System Demonstrations*.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In *Proceedings of IWSLT*.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a simultaneous translation corpus for comparative analysis. In *Proceedings of LREC*.

Katsuhito Sudoh, Takatomo Kano, Sashi Novitasari, Tomoya Yanagita, Sakriani Sakti, and Satoshi Nakamura. 2020. Simultaneous speech-to-speech translation system with neural incremental asr, mt, and TTS. *CoRR*, abs/2011.04845.

Hitomi Tohyama and Yasuyoshi Inagaki. 2004. Ciair simultaneous interpretation corpus.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLLING*.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun HeHua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. Bstc: A large-scale chinese-english speech translation dataset. *arXiv preprint arXiv:2104.03575*.

A Appendix

A.1 Corpus Construction

In this section, we will describe the process of collecting genuine data, and creating interpretation datasets with a proposed dynamic programming algorithm. We will then present a test set annotated by a bilingual annotator to ensure the genuineness of our experimental results and analysis.

A.1.1 Data Collection

We collected genuine data from the European Parliament (EP) Plenary where debates are carried out among representatives of member states of the European Union who speak their native languages; to facilitate communication, simultaneous translation services are provided. From 2008/09/01 to 2012/11/22, source speeches transcriptions, post-edited offline translations and interpretation audios are available. We selected German-English as our target language pair. We used a number of heuristics methods to identify the nationality of the speakers and crawled the data accordingly. Note post-edited translations are only available during this period due to the changes of EP's policies, while audios and transcriptions are provided from 2004 till today. We will leave collecting and making use of the full data in our future work. In total, we downloaded 238 debates and 2415 video files in the mp4 format, with a total size of 500GB. We then used google speech API to transcribe and normalize ASR outputs, while using offline translations to retrieve name entities. Thus, 19,368.24 minutes were transcribed, with a total cost of 697.257 USD.

A.1.2 Data Cleaning

To build a high-quality dataset, we enacted a series of sophisticated automatic and manual preprocessing steps to filter and clean dialogues. The total number of dialogues is 5,239. Initially, we removed dialogues whose source was non-German. To keep the essence of the debates, we filtered out below non-essential components and considered them as transitions between conversations: i) dialogues with less than 20 tokens; ii) dialogues which have pre-defined signals (e.g., "the vote will take place") and they have less than 100-150 words, depending on the signals. The number of data points removed is 2,174. Human investigation on disregarded dialogues confirmed these heuristics. We also discarded those whose source and translation have a different number of sentences, which was, however, rare. The above steps yielded a number of 1,872 data points. Following that, we manually deleted dialogues which were non-essential contents of the debates, while truncating transcriptions whose first and last sentences did not match the corresponding post-edited translation; most of deleted sentences were results of imperfect audio segmen-

Source
• Herr Prsident, liebe Kolleginnen und Kollegen! Wir haben im Ausschuss ber diesen Antrag lange debattiert, wir haben mit grofer Mehrheit eine Entscheidung getroffen, aber es hat gestern und heute eine Flle von Hinweisen und Anregungen gegeben, die sich vor allem auch deshalb ergeben haben, weil andere Ausschsse noch Beratungsgegenstnde hinzugefgt haben.
• Es scheint mir sinnvoll zu sein, nicht heute zu entscheiden, sondern noch einmal die Gelegenheit zu haben, eine Lsung zu finden, die dann auch das Parlament tragen kann. Deshalb bitte ich darum, die Verschiebung heute zu beschlieen. Danke.
Translation
• Mr President, ladies and gentlemen, we debated this motion long and hard in the committee, and we reached a decision backed by a large majority, but yesterday and today, there has been an abundance of advice and suggestions that have come about primarily because other committees have added extra subjects for discussion.
• It seems to me that it would be a good idea not to make the decision today but, instead, to have the opportunity at a later date to find a solution which Parliament is then in a position to support. I therefore ask that you adopt this deferral today. Thank you.
Transcript
• Mr. President dear colleagues in the committee, we discussed this motion at some length.
• We took a decision by a large majority, but between yesterday and today there have been a number of suggestions and indications that have Arisen because other committees. I've also been involved in this procedure.
• So we think it would it would be more intelligent not to take a decision on this report today but to give more time for us to try to find a solution to all of these issues that have been raised that all of Parliament can support. This is why I request that we decide on postponement today.
• Thank you .

Table 5: Example of the constructed dialogues.

tation. After the manual process, 987 dialogues are retained, representing the essence of the debates. Table 5 is one example of the dialogues.

A.1.3 Parallel Dataset Creation

The procedure for constructing parallel interpretation data is described as follows. Firstly, we aligned sentences in the offline translation with those in the interpretation transcription. As the transcription may not be well segmented during the ASR process, we identified sentences in the transcript with stanza[11]. Note each source sentence and post-edited translation sentence in the collected dialogues may comprise several sentences. We call them super-sentences and we don't perform sentence splitting on those super-sentences yet. Next, for each dialogue, we segmented the transcription sentences based on the number of super-sentences in the corresponding translation using dynamic programming, details of which will be discussed in the following section. This step is important due to the fact that unlike post-edited translations which tend to be long and formal, in real-life SI scenarios a source sentence (i.e., German in our case) can often broken into multiple smaller pieces. Hence, it is necessary to recognize and rejoin those pieces into chunks. We chose the candidate, i.e., segmented sentences in the interpretation transcription, which had the highest similarity score to the English translation in the semantic space[12], as the output. More specifically, each of the chunks was semantically similar to one super-sentence in the translation. Since such a super-sentence in any dialogue corresponds to a long, formal German super-sentence, equally each of the chunks can be allocated to that source sentence. This gives us a super-sentence-level dataset, named Super, consisting of 3,683 triples <source, translation, transcript> . This step is done on English pairs, as we believe calculating similarity scores in the same language yields more accurate results than comparing the semantic similarity between different languages.

Following that, we also tried to segment super-sentences with almost the identical procedure[13]. This gave us a sentence-level corpus, the filtered version of which is Clean in the main paper. After inspecting the outputs of the resulting dataset, we noticed it contained noises that were contributed by many factors, the most important of which of is occasional interpreting failure. Hence, we decided to recruit a bilingual German-English speaker to pair sentences manually and they become the test data in this work.

Algorithm 1 Constrained segmentation

Input: $\boldsymbol{Y}$: List of unsegmented target utterances, N : Length of target sequence, $\boldsymbol{X}$: List of segmented source, K : Number of source segments, d: A distance similarity metric.

Output: T: The DP table with optimal scores

// initialisation
for $i = 1...N$ **do**
  $T_{1,i} = d(\boldsymbol{X}_1, \boldsymbol{Y}_{1:i})$
end for
// Filling out T based on the DP relation
for $k = 2...K$ **do**
  for $i = k...N$ **do**
    $T_{k,i} = \max_{k \leq j \leq i} [T_{k-1,j-1} + d(\boldsymbol{X}_k, \boldsymbol{Y}_{j:i})]$
  end for
end for
return T

A.1.4 Sequence Segmentation/Alignment with Dynamic Programming

We use a dynamic programming algorithm, as shown in Algorithm 1, to segment target utterances and perform alignment in the semantic space. As shown in Table 5, each source sentence has its own correspondence of translation, so we only need to align segments of sentences to that translation sentence, in order to have a parallel source-interpretation corpus. Hence, we dynamically divide sentences in the transcription by the number of translation sentences, calculate the similarity score for each pair while considering the accumulated scores for sequences preceding it. We then trace-back the candidate with the best score. The time complexity of this algorithm is $O(KN^2)$, where K is the number of source sentences and N is the number of target utterances. This algorithm is applicable to creating both the super-sentence-level and sentence-level parallel datasets.

A.2 Europarl vs Style Transferred Europarl

To illustrate the outcomes of style transfer models, we provides illustrations of target sentences in Europarl and style transferred Europarl (Pseudo-I) in

[11] https://github.com/stanfordnlp/stanza

[12] Similarity scores are calculated with https://github.com/UKPLab/sentence-transformers/tree/master/sentence_transformers

[13] The only difference here is that we used a multilingual encoder

Europarl	Style Transferred Europarl
There has therefore been enough time for the Commission to prepare its programme and for us to become familiar with it and explain it to our citizens.	So there has been enough time for the Commission to draw up the program and for us to be aware that and explain it to our citizens.
I would urge you to endorse this.	I would ask you to agree with that.

Table 6: Examples of Europarl vs Style Transferred Europarl.

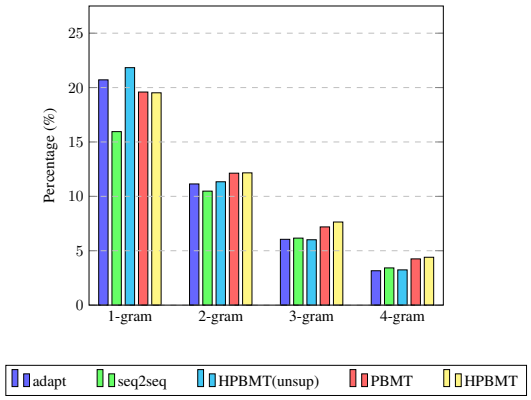

Figure 2: Percentage of introduced correct n-grams.

Table 6. The first example involves word reordering and word replacement that "so" is replaced with "therefore" before being put at front; it also involving word replacement in that "prepare its programme" is changed to "draw up the program" and "become familiar with it" changed to "be aware that". In the second example, "endorse" is replace with a common phrase "agree with".

A.3 Discussion

A.3.1 Analysis on N-grams.

To investigate what led to the improvement, we first computed n-grams present in gold Interpretation but not in outputs predicted by the baseline wait-k. Then we examined the amount of n-grams newly introduced by each model that are overlapped with the n-grams calculated previously. As shown[14] in Figure 2, PBMT and HPBMT in supervised settings consistently introduce more n-grams than others with the only exception that adapt and HPBMT(unsup) produce more new 1-gram. Essentially, this implies that style transferred Europarl has effectively captured more interpretation features than the original Europarl.

[14]We dropped unsupervised, supervised for the sake of clarity of the plot, while just using HPBMT(unsup) to indicate HPBMT(unsupervised).