Knowledge Distillation Improves Stability in Retranslation-based Simultaneous Translation

Anonymous ACL submission

Abstract

In simultaneous translation, the *retranslation* approach has the advantage of requiring no modifications to the inference engine. However in order to reduce the undesirable instability (flicker) in the output, previous work has resorted to increasing the latency through masking, and introducing specialised inference, losing the simplicity of the approach. In this paper, we argue that the flicker is caused by both non-monotonicity of the training data, and by non-determinism of the resulting model. Both of these can be addressed using knowledge distillation. We evaluate our approach using simultaneously interpreted test sets for English-German and English-Czech and demonstrate that the distilled models have an improved flicker-latency tradeoff, with quality similar to the original.

1 Introduction

Simultaneous machine translation systems, which process their input word by word instead of sentence by sentence, must strike a balance between producing output immediately (and so reducing quality because of incomplete input) and waiting for further input (and so increasing latency). A good simultaneous translation system will provide a pareto-optimal tradeoff between quality and latency. A straightforward way of doing simultaneous translation is *retranslation* (Niehues et al., 2016), which has the advantage that it can be used with an unmodified machine translation (MT) inference engine, and can perform better than the alternative, streaming-based approaches (Arivazhagan et al., 2020b). The disadvantage is that retranslation may change previous output causing *flicker*, leading to a poor user experience, and needs to be balanced with latency and quality.

We argue that flickering is caused by two different (but related) issues: (i) instability of the translation – the system "changes its mind" as more source is revealed; (ii) non-monotonicity of the translation – the system favours a non-monotonic translation, which means it needs high latency in order to avoid flicker. Some of this instability and non-monotonicity is necessary – forced by syntactic differences between source and target, and lack of information in the prefixes – but some is due to arbitrary choices of the model and we aim to reduce these as much as possible.

Researchers in non-autogressive translation (NAT) have identified a related problem, known as the "multimodality" problem (Gu et al., 2018), where the model has two or more high scoring translations but outputs a poor quality mixture of them (because of the independence assumptions in NAT). The solution to this problem is to use sequence-level knowledge distillation (Kim and Rush, 2016), which was also shown to result in more monotonic translations (Zhou et al., 2020). In simultaneous translation, we observe a different type of multimodality (see Table 4), where the model has two competing translations (which may be synonyms) and flips between the two, unnecessarily. We therefore investigate whether the same solution as proposed there, i.e. knowledge distillation or teacher-student models, can also reduce flicker in simultaneous translation. We will show that an appropriately trained student model, in other words a model trained on a synthetic corpus created by translating using a teacher model, is able to achieve the same quality as the teacher, but with substantially lower flicker.

2 Background

We focus on simultaneous translation using the retranslation approach, and in particular how to stabilise the output, without reducing quality, and without sacrificing the simplicity of the inference.

The problem of reducing flicker was considered by Arivazhagan et al. (2020a), who showed that masking the last k words of the output, combined with biasing the beam search towards the previously translated prefix could improve the flicker-latency tradeoff, although this required modifications to the inference engine. To set the mask dynamically, Yao and Haddow (2020) showed that the system could make predictions of the continuation of the prefix, and compare the translations of these continuations to the translations of the current prefix. However this method has the disadvantage of requiring extra translation inference, making it less efficient at runtime.

Evaluation of simultaneous translation requires that we consider more than just the quality of translation, we must also consider the latency, and if we are using retranslation, we should consider flicker. The quality of the translation can evaluated by comparing the final output of each sentence with a reference – we will show BLEU (Papineni et al., 2002; Post, 2018), CHRF (Popovi, 2015) and COMET (Rei et al., 2020) scores. For evaluation of flicker, we will use *normalised erasure* (Arivazhagan et al., 2020a), which measures the number of tokens that must be deleted from the suffix of the previous translation to produce the next, normalised by sentence length. The measurement of latency has been the subject of some debate in the literature, with several different measures proposed (Ma et al., 2019a; Cherry and Foster, 2019; Ansari et al., 2021), and for retranslation systems there is the further question of whether to use the time that a word appears, or the time that it stabilises, in the latency calculation. In our experiments, we will vary the amount of output masking, and observe the effect on flicker. The amount of masking is a clear measure of how much delay there is in the translation, and is easily controllable. The aim is to improve the mask-flicker tradeoff curve, and so be able to use a shorter mask with the same flicker budget.

In sequence-level knowledge distillation (Kim and Rush, 2016), a smaller *student* model is created using data generated by the larger *teacher* model. This has found application in MT efficiency (Junczys-Dowmunt et al., 2018), where the small size of the student models ensure that they make inference much faster, and they can also be run using a small beam. In non-autoregressive translation, teacher-student models are able to reduce the multimodality problem – by reducing the number of possible translations favoured by the model, the effect of the conditional independence assumption in NAT is mitigated (Zhou et al., 2020).

For our purposes, teacher-student methods play a similar role. Because the student model tends to prefer a single translation hypothesis, the model is less likely to swap between translation hypotheses unnecessarily as the source prefix is extended. Also, since the student model is trained on MT output, where the target order tends to be similar to the source order, the student is more likely to avoid unnecessary reorderings, generating a more monotone translation, which can be built up incrementally. We will demonstrate these points experimentally in the next section.

Recently, Chen et al. (2021) also proposed to use pseudo-reference sentences obtained through forward translation of the source sentences to improve simultaneous translation. Unlike our work, they considered a streaming approach (specifically wait-k (Ma et al., 2019b)) where the system can only append to the output, it does not flicker like retranslation. They showed that they could improve the quality-latency tradeoff of wait-k using their distillation approach, but to create the training data for the student system they used wait-k and filtering – we avoid these complications by just using the baseline system as the teacher.

3 Experiments

3.1 Data

In much of the previous work on simultaneous MT, models are evaluated on translations that were produced offline, where the translators could access the full sentence. As pointed out by Zhao et al. (2021), this may not be a realistic evaluation. So in this work, we test on the recently released ESIC corpus (Macháek et al., 2021), a corpus derived from the European parliament proceedings which contains both transcripts of the original speeches, and transcripts of the simultaneous interpretation of those speeches. ESIC also contains the corresponding text-based records, which can be considered as offline translations. ESIC is available for English→Czech and English→German, and it is aligned at the document level, but not at the sentence level. We use the test portion for evaluation.

We train our systems using offline translations, as there are no large corpora of simultaneous interpretation for training. For English→German, we use the IWSLT 2021 data sets (Anastasopoulos et al., 2021). This includes the English→German data from WMT 2020 (Barrault et al., 2020). For development, we use the concatenation of IWSLT test sets from 2014 and 2015. We removed the train/test overlaps – between MuST-C.v2 and earlier IWSLT test sets, and between europarl and ESIC. For English→Czech, we use the training and valid set from WMT21 (Akhbardeh et al., 2021). Training data sizes are shown in Table 3.

3.2 Teacher System

Our initial system, which will later be used as a teacher model (Section 3.3), is a transformer base model[1] (Vaswani et al., 2017) trained with marian (Junczys-Dowmunt et al., 2018). We use *prefix training* to reduce the mismatch between sentence-level training data and prefix-based inference at test time (Niehues et al., 2018). For each parallel sentence pair in the training set, we generate a corresponding prefix pair by truncating using a randomly chosen proportionate length.

All data is pre-processed using a unigram language model (Kudo, 2018) with SentencePiece (Kudo and Richardson, 2018) with a shared subword (Sennrich et al., 2016) vocabulary size of 32k. We train the MT models to convergence (using early stopping of 10) with a learning rate of 0.0003, and translate using a beam of 6.

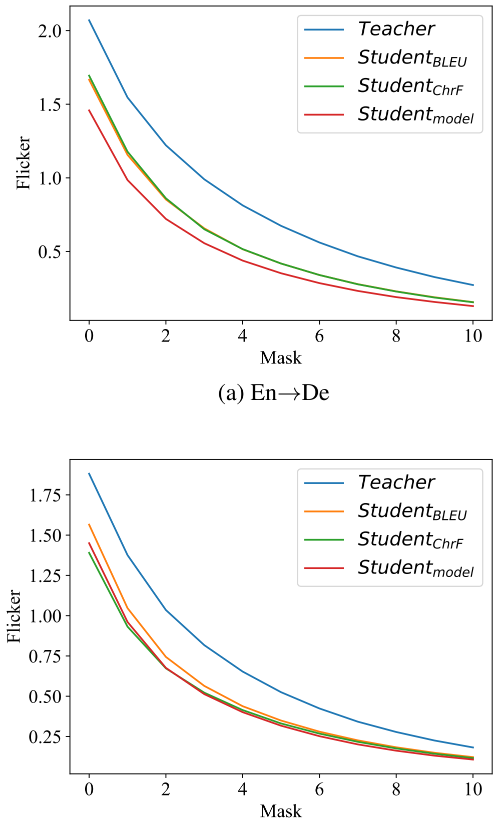
3.3 Teacher-Student Training

In order to create a more stable system, we use the teacher model in the previous section to generate training data for student models. These student models are trained in the same way, with the same architecture, but with training data synthesised by the teacher. For each source sentence, we generate n-best translations and then select the best translation that has highest score against the reference translation. In our experiments we consider 8-best translation. We use three different scores (BLEU, CHRF, and model[2] score), to select distilled training data.

In order to calculate the monotonicity of the training data, we use Kendall's tau distance. To compute the distance, we first align the parallel data using *fast_align* (Dyer et al., 2013) and then find the source permutation π of a target sentence

[1] With 65 million parameters.

[2] For distillation using model score, we do not compare with a reference translation. Instead, each source is forward translated into the target language by the teacher model and we take the highest scoring translation.



(a) En→De

(b) En→Cs

Figure 1: Sentence level Flicker vs Latency plot. The y-axis represents flicker and the x-axis represents the number of words that are masked.

as

$$\pi = \{j : i^{th} \text{ target word is aligned to } j^{th} \text{ source word}\}$$

We calculate the Kendall's tau distance between π and $\pi^{'}$, where

$$\pi^{'} = \{i : i^{th} \text{ target word }\}$$

The scores are calculated at the sentence level and then averaged over a parallel corpus. The higher tau score indicates more monotonicity.

In our experiments, we find the distance between

- the source and reference (Source-Reference)
- the source and 1-best distilled target (Source-Distilled$_{\text{model}}$)
- the source and distilled target obtained from n-best using BLEU score (Source-Distilled$_{\text{BLEU}}$)
- the source and distilled target obtained from n-best using ChrF score (Source-Distilled$_{\text{ChrF}}$)

	Model	BLEU	ChrF	COMET-qe	Flicker
		En→De			
Interpreted	Teacher	17.6	59.0	0.539	2.07
	$\text{Student}_{\text{model}}$	17.5	58.9	0.530	1.46 (29.46% ↓)
	$\text{Student}_{\text{BLEU}}$	17.6	58.9	0.527	1.67 (19.32% ↓)
	$\text{Student}_{\text{ChrF}}$	17.6	59.0	0.530	1.69 (18.35% ↓)
		En→Cs			
	Teacher	14.6	51.7	0.680	1.88
	$\text{Student}_{\text{model}}$	14.6	51.7	0.660	1.45 (22.87% ↓)
	$\text{Student}_{\text{BLEU}}$	14.6	51.7	0.670	1.56 (17.02% ↓)
	$\text{Student}_{\text{ChrF}}$	14.7	51.8	0.661	1.39 (26.06% ↓)
		En→De			
Translated	Teacher	36.4	63.7	0.540	2.61
	$\text{Student}_{\text{model}}$	36.0	63.4	0.533	1.70 (34.86% ↓)
	$\text{Student}_{\text{BLEU}}$	36.4	63.6	0.534	1.94 (25.67% ↓)
	$\text{Student}_{\text{ChrF}}$	36.6	63.9	0.532	2.02 (22.60% ↓)
		En→Cs			
	Teacher	33.9	60.0	0.721	2.33
	$\text{Student}_{\text{model}}$	33.3	59.7	0.693	1.62 (30.47% ↓)
	$\text{Student}_{\text{BLEU}}$	33.9	60.1	0.701	1.81 (22.31% ↓)
	$\text{Student}_{\text{ChrF}}$	34.0	60.2	0.694	1.66 (28.75% ↓)

Table 1: Comparison between different approaches on ESIC test set. BLEU and ChrF scores are calculated at document level for Interpreted category and at sentence level for translated category using Sacrebleu. The COMET-qe score is calculated between source and the hypothesis using reference-less *wmt20-comet-qe-da* model. We use reference-less scoring as we do not have equal number source and reference lines for interpreted ESIC corpus. The flicker scores are calculated at sentence level on outputs without any mask. In parentheses, we show relative reduction in flicker.

Model	Pair	Distance
En→De	Source-Reference	0.793
	$\text{Source-Distilled}_{\text{BLEU}}$	0.826
	$\text{Source-Distilled}_{\text{ChrF}}$	0.848
	$\text{Source-Distilled}_{\text{model}}$	0.857
En→Cs	Source-Reference	0.849
	$\text{Source-Distilled}_{\text{BLEU}}$	0.900
	$\text{Source-Distilled}_{\text{ChrF}}$	0.904
	$\text{Source-Distilled}_{\text{model}}$	0.906

Table 2: Kendall's tau distances. Higher scores indicate more monotonicity.

We have presented the tau scores in Table 2. From Table 2, we observe that the distillation makes the training data more monotonic and 1-best distilled data has the best tau distance.[3]

3.4 Stability of Student Models

We calculate the BLEU score at sentence and document level using Sacrebleu for translated and interpreted ESIC testset, respectively, and flicker at sentence level using SLTev toolkit (Ansari et al., 2021). We compare the quality of teacher and student models in Table 1.

We observe that student models have a substantially reduced flicker (by 17-34%) with no loss in either document or sentence-level BLEU or ChrF scores, although there is a moderate drop in COMET-qe. The flicker can be further reduced with masking the subsequent output prefixes. We apply different fixed mask of length 1-10 and plot the flicker (measure using normalized erasure) against each fixed mask in Figure 1. Masking helps reducing the flicker and the student models flicker less than the teacher for a given mask length. Since quality is calculated on the final output, masking does not impact BLEU/chrF/COMET.

4 Conclusion

In this paper, we proposed to reduce the flicker in retranslation-based simultaneous translation through knowledge distillation. We use different metrics to select the synthetic target-side data, which are monotonic measured using Kendall's tau distance, from n-best forward translations. We use the synthetic data to train the retranslation-based simultaneous translation system. Our evaluation on interpreted testsets for English-German and English-Czech show significant reduction in the flicker with similar quality as the teacher.

[3] Additionally, we use tau distance to filter the 1-best distilled data, and then we train more models on the filtered data. For filtering purpose, we sort the distilled parallel corpus by monotonicity and take top 90, 80, 70, and 60% parallel sentences for training student models. But this did not reduce the flicker further significantly.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–93, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. Findings of the IWSLT 2021 Evaluation Campaign. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTEV: Comprehensive evaluation of spoken language translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Te I, Wolfgang Macherey, Pallavi Baljekar, and George Foster. 2020a. Re-Translation Strategies For Long Form, Simultaneous, Spoken Language Translation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020b. Re-translation versus Streaming for Simultaneous Translation. ArXiv: 2004.03643v2.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Junkun Chen, Renjie Zheng, Atsuhito Kita, Mingbo Ma, and Liang Huang. 2021. Improving simultaneous translation by incorporating pseudo-references with fewer reorderings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5857–5864, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Cherry and George Foster. 2019. Thinking Slow about Latency Evaluation for Simultaneous Machine Translation. ArXiv: 1906.00048v1.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-Autoregressive Neural Machine Translation.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective High-Quality Neural Machine Translation in C++. In *Proceedings of WNMT*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327. Association for Computational Linguistics. Event-place: Austin, Texas.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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. 2019a. 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.

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. 2019b. 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 Macháek, Matú ilinec, and Ondej Bojar. 2021. Lost in Interpreting: Speech Translation from Source or Interpreter? In *Proceedings of Interspeech*.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic Transcription for Low-latency Speech Translation. In *Proceedings of Interspeech*.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *Proceedings of Interspeech*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. Type: Conference proceedings (article).

Maja Popovi. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.

Yuekun Yao and Barry Haddow. 2020. Dynamic Masking for Improved Stability in Online Spoken Language Translation. In *Proceedings of AMTA*.

Jinming Zhao, Philip Arthur, Gholamreza Haffari, Trevor Cohn, and Ehsan Shareghi. 2021. It is Not as Good as You Think! Evaluating Simultaneous Machine Translation on Interpretation Data. *arXiv:2110.05213 [cs]*. ArXiv: 2110.05213.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Appendix

Corpus	Sentence pairs
English-German	
Europarl	1.79 M
Rapid	1.45 M
News Commentary	0.35 M
OpenSubtitle	22.51 M
TED corpus	206 K
MuST-C.v2	248 K
English-Czech	
Europarl	645 K
ParaCrawl	14 M
CommonCrawl	161 K
News Commentary	260 K
CzEng2.0	36 M[4]
Wikititles	410 K
Rapid	452 K

Table 3: Corpora used in training the systems

Source	I hope you will have a little time and energy to focus on another report which is, despite its technicality, quite important for all of us.
Traget:	Ich Ich hoffe, Ich hoffe, Sie Ich hoffe, Sie Ich hoffe, Sie haben Ich hoffe, Sie haben ein Ich hoffe, Sie werden ein wenig Zeit Ich hoffe, Sie haben etwas Zeit Ich hoffe, Sie haben etwas Zeit und Ich hoffe, Sie werden etwas Zeit und Energie haben, Ich hoffe, Sie haben etwas Zeit und Energie, um sich Ich hoffe, Sie haben etwas Zeit und Energie, um sich auf Ich hoffe, Sie werden ein wenig Zeit und Energie haben, um sich auf ein anderes Thema Ich hoffe, Sie haben etwas Zeit und Energie, um sich auf einen weiteren Bericht zu konzentrieren, Ich hoffe, Sie haben etwas Zeit und Energie, um sich auf einen anderen Bericht zu konzentrieren, ⋮ Ich hoffe, Sie werden ein wenig Zeit und Energie haben, um sich auf einen anderen Bericht zu konzentrieren, der trotz seiner Formalität für uns alle sehr wichtig ist.

Table 4: Examples of flicker caused by the teacher model. *Source* is the original full sentence which is input as a growing input prefix. *Target* is the output prefix in successive retranslations.