On the Hallucination in Simultaneous Machine Translation

Meizhi Zhong[1], Kehai Chen[1] *, **Zhengshan Xue[2], Lemao Liu, Mingming Yang, Min Zhang[1]**

[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
22s051052@stu.hit.edu.cn, chenkehai@hit.edu.cn, xuezhengshan@tju.edu.cn,
lemaoliu@gmail.com, shanemmyang@gmail.com, zhangmin2021@hit.edu.cn

Abstract

It is widely known that hallucination is a critical issue in Simultaneous Machine Translation (SiMT) due to the absence of source-side information. While many efforts have been made to enhance performance for SiMT, few of them attempt to understand and analyze hallucination in SiMT. Therefore, we conduct a comprehensive analysis of hallucination in SiMT from two perspectives: understanding the distribution of hallucination words and the target-side context usage of them. Intensive experiments demonstrate some valuable findings and particularly show that it is possible to alleviate hallucination by decreasing the over usage of target-side information for SiMT. [1]

1 Introduction

In neural machine translation, hallucination occurrences are not common due to its small quantity (Lee et al., 2018; Yan et al., 2022; Raunak et al., 2021a; Guerreiro et al., 2023). But in simultaneous machine translation (SiMT), it has been found that hallucination is extremely severe, especially as latency increases indicating that hallucination is a critical issue in SiMT. Currently, most prior works concentrate on how to enhance model performance for SiMT (Ma et al., 2019, 2020; Zheng et al., 2020; Zhang and Feng, 2022a,b; Guo et al., 2022; Zhang and Feng, 2022c), however, only a few of them measure the hallucination phenomenon (Chen et al., 2021; Deng et al., 2022; Liu et al., 2023). To our best knowledge, there are no researches which *systematically analyze hallucination in SiMT*.

Therefore, we conduct a comprehensive analysis of hallucinations in SiMT. Initially, we seek to empirically analyze these hallucination words from

the perspective of their distribution. We collect all hallucination words together and understand their frequency distribution, and we find that these words are randomly distributed with a high entropy: their entropy is almost as high as that for all target words. In addition, to delve into the contextual aspects of hallucination (Xiao and Wang, 2021), we consider their predictive distribution. We discover that their uncertainty is significantly higher than that of non-hallucination words. Furthermore, we find that the SiMT model does not fit the training data well for hallucination words due to the essence of SiMT (i.e., the limited source context), which explains why making correct predictions for hallucination words is difficult.

Intuitively, since a SiMT model is defined on top of a limited source context, this may indirectly cause the model to focus more on the target context and lead to the emergence of hallucination words. To verify this intuition, we propose to analyze the usage of the target context for hallucination words for SiMT. Specifically, following Li et al. (2019); Miao et al. (2021); Fernandes et al. (2021); Voita et al. (2021); Yu et al. (2023); Guerreiro et al. (2023), we firstly employ a metric to measure how much target-context information is used by SiMT with respect to the source-context information. With the help of this metric, we find that hallucination is indeed significantly more severe when the SiMT model focuses more on target-side information. Drawing upon this, we reduce the over-target-reliance effects by introducing noise into the target-side context. Experimental results show that the proposed method achieves some modest improvements in terms of BLEU and hallucination effect when the latency is relatively small. This discovery gives us some inspiration: more flexible control over the use of target-side information may be a promising approach to alleviate the issue of hallucination.

Our key contributions are as follows:

- We study hallucination words from frequency and predictive distributions and observe that the frequency distribution of hallucination words is with high entropy and hallucination words are difficult to be memorized by the predictive distribution during training.

- We analyze hallucination words according to the usage of (limited) source context. We find that hallucination words make use of more target-context information than source-context information, and it is possible to alleviate hallucination by decreasing the usage of the target context.

2 Experimental Settings

Our analysis is based on the most widely used SiMT models and datasets. This section introduces these models and datasets as follows.

SiMT Models and Datasets. SiMT models translate by reading partial source sentences. Ma et al. (2019) proposed widely used Wait-k models for SiMT. It involves reading k words initially and then iteratively generating each word until the end of the sentence. We conducted experiments on it. We use two standard benchmarks from IWSLT14 De↔En (Cettolo et al., 2013) and MuST-C Release V2.0 Zh→En (Cattoni et al., 2021) to conduct experiments. Appendix A provides detailed settings. Due to space limitation, we only present the experimental results for the De→En benchmark. The results for Zh→En and En→De are similar, as shown in Appendix D and C.

Hallucination Metric. In SiMT, Chen et al. (2021) pioneers the definition of Hallucination Metrics based on word alignment a. A target word $\hat{y}_t$, is a hallucination if there is no alignment to any source word x_j. This is formally represented as:

$$H(t, a) = \mathbb{1}\left[\{(i, t) \in a\} = \varnothing\right]. \quad (1)$$

Conversely, a target word $\hat{y}_t$, is not a hallucination if there is alignment to any source word x_j.

The Hallucination Rate (HR) is defined as following:

$$\mathrm{HR}(x, \hat{y}, a) = \frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} H(t, a). \quad (2)$$

Deng et al. (2022) propose GHall to measure hallucination in Wait-k. Formally, a word is a

k	1	3	5	7	9	∞
HR %	31.28	22.57	18.58	16.41	15.21	11.50

Table 1: HR on valid set of wait-k, where $k = \infty$ means Full-sentence MT.
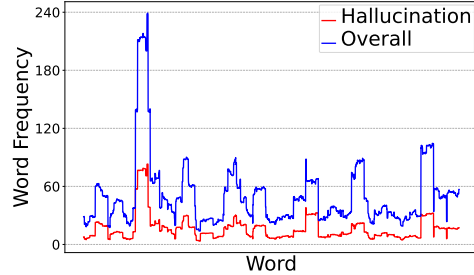

Figure 1: Word frequency of Hallucination and Overall on valid hypotheses set of wait-1 (x-axis is ordered randomly, with additional k results in Appendix B.1).

hallucination if it does not align with the current source:

$$H_{wait-k}(t, a) = \mathbb{1}[\{(s, t) \in a \mid s \geq t + k\} = \varnothing]. \quad (3)$$

The definition of HR remains consistent with Chen et al. (2021). We utilize GHall metrics to conduct experiments. We use Awesome-align (Dou and Neubig, 2021) as the word aligner a.

3 Understanding Hallucination Words from Distribution

Hallucination is severe in SiMT. We measure HR of Wait-k models, illustrated in Table 11. We obtain that Wait-k models suffer more from hallucinations than Full-sentence MT. Furthermore, with k decreasing, hallucinations increase clearly. This shows that hallucination is an important issue and it is worth the in-depth study.

3.1 Understanding Hallucination from Frequency Distribution

Hallucination words are with high distribution entropy. To investigate hallucination words in Wait-k, we compare frequency distributions of hallucination and overall words. Figure 1 and Table 2 illustrate that their distributions are remarkably similar and both exhibit high entropy. It suggests that understanding hallucination from high distribution entropy is challenging.

k	1	3	5	7	9
Hallucination	7.82	8.22	8.19	8.10	8.07
Overall	8.70	8.97	9.00	9.01	9.02

Table 2: Word frequency distribution entropy of Hallucination and Overall on the valid set of wait-k.

Wait-k	Valid set				Training subset			
	Uncertainty		Confidence		Uncertainty		Confidence	
	H	NH	H	NH	H	NH	H	NH
$k=1$	3.53	2.35	0.40	0.61	3.47	2.13	0.41	0.65
$k=3$	3.00	2.04	0.48	0.66	2.98	1.90	0.49	0.69
$k=5$	2.81	1.97	0.52	0.67	2.76	1.90	0.52	0.69
$k=7$	2.55	1.89	0.55	0.69	2.48	1.81	0.57	0.70
$k=9$	2.48	1.92	0.57	0.68	2.42	1.96	0.58	0.69

Table 3: The Uncertainty and Confidence of Hallucination (**H**) and Non-Hallucination (**NH**) on the valid set and training subset of wait-k models.

3.2 Understanding Hallucination from Predictive Distribution

We investigate **Confidence** and **Uncertainty** of the predictive distribution. We define the Confidence of a word as its probability and the Uncertainty of a word as the entropy of its predictive distribution.

Hallucination words are difficult to translate. To explore the difficulty of translating hallucination and non-hallucination words, we calculate the average confidence and uncertainty on the valid set. The results in the left of Table 3 reveal that during decoding hallucination words, the models exhibit higher uncertainty. Additionally, the confidence is lower. It suggests that models encounter challenges in accurately translating hallucination words.

Hallucination words are difficult to memorize. To investigate the reasons behind the difficulty in translating hallucination words, we measure confidence and uncertainty for hallucination and non-hallucination words on the training data. We sample examples from the training data as a training subset with the same size as the valid set. The results in the right of Table 3 illustrate that even in previously encountered contexts, models remain uncertain when dealing with hallucination words. These findings suggest that models do not fit well with hallucination words during training, leading to a limited ability to generalize to similar contexts on the valid set. Consequently, the difficulty in translating hallucination words can be attributed to challenges in memorization during the training. Additionally, we observe that as k increases, the

uncertainty decreases significantly. It can be attributed to the model encountering source-side context more, enabling a improved memorization.

4 Analysis of Target Context Usage for Hallucination Words

To verify the hypothesis that using more on target-side context leads to the emergence of hallucination, we propose to analyze the usage of target-side context.

Measure on Target-side Context Usage. To explicitly measure Target Context Usage, we adapt an interpretive approach that evaluates the relevance of both target and source words. It involves deactivating connections between the corresponding words and the network. We compute the relevance between the words in the source or target and the next word to be generated and determine the maximum absolute relevance as source or target relevance. It allows us to calculate the **T**arget-Side Relevance to **S**ource-Side Relevance 's **R**atio (TSSR).

To begin with, we assess the relevance of target-side words and source-side words to the next word to be generated. This evaluation is conducted by selectively deactivating the connection between x_j or y_j and the encoder or decoder network in a deterministic manner, following the approach described in Li et al. (2019). More formally, the relevance $R(y_i, x_j)$ or $R(y_i, y_j)$ in Wait-k is directly determined through the dropout effect on x_j or y_j, as outlined below:

$$R(y_i, x_j) = P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}_{\leq i+k-1}) \\ - P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}_{\leq i+k-1,(j,\mathbf{0})}). \quad (4)$$

$$R(y_i, y_j) = P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}_{\leq i+k-1}) \\ - P(y_i \mid \mathbf{y}_{<i,(j,\mathbf{0})}, \mathbf{x}_{\leq i+k-1}). \quad (5)$$

The relevance of the source-side and target-side is determined by selecting the maximum absolute value of the word's relevance on the current source-side and the current target-side. Formally, this can be expressed as:

$$R(y_i)_{source-side} = \max\{|R(y_i, x_j)|\}. \quad (6)$$

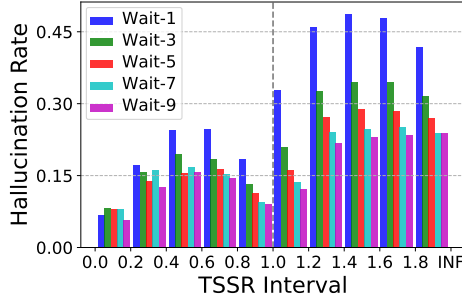$$R(y_i)_{target-side} = \max\{|R(y_i, y_j)|\}. \quad (7)$$

Figure 2: HR on the valid set in different TSSR intervals of wait-k models.

Finally, the ratio of target-side relevance to source-side relevance(TSSR) is calculated. A larger TSSR indicates a higher usage of target-side context in generating the next word y_i.

$$TSSR\left(\mathrm{y}_i\right) = \frac{R\left(\mathrm{y}_i\right)_{target-side}}{R\left(\mathrm{y}_i\right)_{source-side}}. \qquad (8)$$

Our final algorithm, referred to as Algorithm 1, is presented.

Algorithm 1 Compute TSSR

Input: model, hypotheses sentence, source sentence, k
Output: TSSR
for i in hypotheses sentence length **do**
    if $j < i$ **then**
        Compute the relevance of next word y_i and y_j according to 5
    end if
end for
for i in source sentence length **do**
    if $j \le i + k - 1$ **then**
        Compute the relevance of next word y_i and x_j according to 4
    end if
end for
Compute Target-Side Relevance according to 7
Compute Source-Side Relevance according to 6
Compute TSSR according to 8

TSSR is categorized into 10 intervals from 0 to INF, indicating the degree of Target Context Usage.

4.1 The Relationship between Hallucination and Target-side Context Usage

Using more target context leads to more severe hallucination. Initially, we analyze the relationship between a word's usage of the
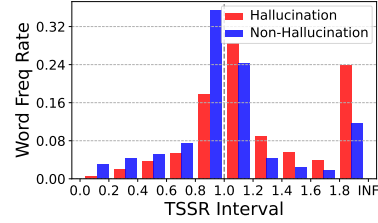
(a) De-En
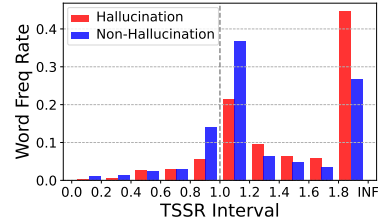
(b) Zh-En (human alignment annotation)

Figure 3: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for wait-1 model.

target-side context and its likelihood of being a hallucination. Building upon this, we explore the HR across different TSSR intervals, as depicted in Figure 2. Our findings demonstrate that in high TSSR intervals, HR is higher compared to low TSSR intervals. It indicates that a word using more target context is more likely to be a hallucination.

Further analysis revealed that when comparing different Wait- values, there is a more pronounced increase in HR from low TSSR intervals to high TSSR intervals as k decreases, as depicted in Figure 2. This means that there maybe an increased likelihood of hallucinations occurring in words that are utilized with limited source-side context

Hallucination words use more target context than Non-Hallucination words. The aforementioned analysis motivates us to investigate whether hallucination words indeed exhibit a higher usage of target-side context than non-hallucination words. To explore this, we analyze the TSSR distributions of hallucination and non-hallucination word frequencies. Figure 3(a) reveals that hallucination words are concentrated on high TSSR intervals. This means the model tends to use more target-side context for the generation of a hallucination word. Furthermore, we observed that the word frequency rate of non-hallucination words is higher in the 0.8 ~1.2 TSSR range, also illustrated in Figure 3(a). Therefore, we propose that the model utilizes source-side context and
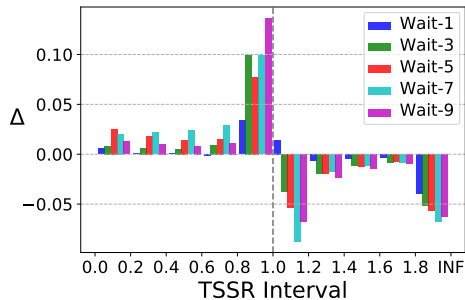
Figure 4: Word Frequency Rate Change (Δ) in different TSSR intervals with scheduled sampling training compared to the Baselines.

Figure 5: Hallucination Frequency Change (Δ) in different TSSR intervals with scheduled sampling training compared to the Baselines.

		$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
Baselines	BLEU ↑	19.69	26.76	29.61	31.10	32.03
	HR % ↓	31.28	22.57	18.58	16.41	15.21
Scheduled-Sampling	BLEU ↑	20.53	27.32	30.23	31.73	32.34
	HR % ↓	30.85	21.62	17.84	15.16	13.84

Table 4: BLEU scores and HR of wait-k models.

target-side context similarly during the generation of non-hallucination words. To further validate our claims of above analysis, we sample 100 sentences from the translation results of Zh-En using wait-1 decoding for human alignment annotation. We then conduct experiments similar to Figure 3(a). The results as shown in Figure 3(b) are consistent with the conclusions drawn in automatic alignment annotation.

4.2 Increasing Source-side Context Usage via Reducing Target-side Context Usage

Observing the association between hallucination and usage of target-side context, we posit that reducing this reliance might be a viable approach to mitigate the hallucination in SiMT. Inspired by (Bengio et al., 2015; Zhang et al., 2019), we adopt the scheduled sampling training to guide the models to pay more attention on the source-side context by adding noise to the target-side context. Specifically, we randomly replace the ground truth tokens with predicted ones using a decaying probability. The results shown in Figure 4 indicate a decrease in target-context usage and an increase in source-context usage. Scheduled sampling training exhibits improvements in BLEU scores and reductions in HR as presented in Table 4. It successfully reduces hallucination words using more target-side context, but also indirectly increases hallucination words using more source-side context, as shown in Figure 5. Therefore, a
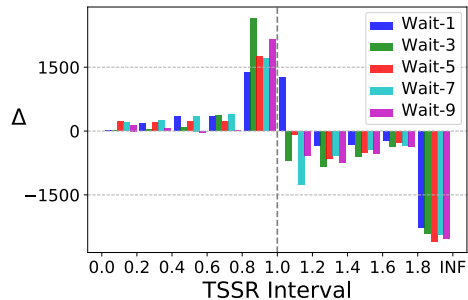
better method to flexibly handle the usage between target-side and source-side context is required.

5 Related Work

In NMT, previous works have delved into the phenomenon of hallucinations(Lee et al., 2018; Müller et al., 2020; Wang and Sennrich, 2020; Raunak et al., 2021b; Zhou et al., 2021). Specifically, Voita et al. (2021) assessed the relative contributions of source and target context to predictions. Weng et al. (2020); Miao et al. (2021) argued that an important reason for hallucination is the model's excessive attention to partial translations in NMT. Furthermore, Guerreiro et al. (2023) conducted a comprehensive study of hallucinations in NMT. Differing from these works focusing on NMT, this paper conducts a comprehensive analysis of hallucination in SiMT.

6 Conclusions

This paper conducts the first comprehensive analysis of hallucinations in SiMT from two perspectives: understanding the hallucination words from both frequency and predictive distributions and their effects on the usage of target-context information. Intensive Experiments demonstrate some valuable findings: 1) the frequency distribution of hallucination words is with high entropy and their predictive distribution is with high uncertainty due to the difficulty in memorizing hallucination words during training. 2) hallucination words make use of more target-side context than source-side context, and it is possible to alleviate hallucination by decreasing the usage of target-side context.

Limitations

We highlight four main limitations of our work.

Firstly, instead of focusing on more recent adaptive policy, our analysis focuses on the hallucinations in the Wait-k Policy (Ma et al., 2019), which is the most widely used fixed policy in SiMT to ensure a simple and familiar setup that is easy to reproduce and generalize.

Secondly, although we propose a simple methods to control the usage of target information, attempting to mitigate the hallucination in SiMT, we only achieve limited improvement. In the future, we will explore more flexible and robust approaches for controlling target context usage to better mitigate the hallucination and achieve greater performance.

A further limitation of our study is that we exclusively analyze hallucinations as defined in Section 2, without considering detached hallucinations. This omission arises from the absence of established and reliable automated evaluation methods for detecting such detached hallucinated words.

Moreover, our study is constrained by its reliance on aligner tools, potentially introducing alignment biases. Therefore, when applying our approach to datasets with lower alignment accuracy, careful consideration is warranted regarding the necessity for additional validation and adjustment.

Acknowledgements

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*, Heidelberg, Germany.

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.

Hexuan Deng, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. 2022. Improving Simultaneous Machine Translation with Monolingual Data.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient wait-k models for simultaneous machine translation. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1461–1465. ISCA.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. 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 6467–6478, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Shoutao Guo, Shaolei Zhang, and Yang Feng. 2022. Turning fixed to adaptive: Integrating post-evaluation into simultaneous machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2264–2278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *NIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop*.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

Mengge Liu, Wen Zhang, Xiang Li, Yanzhi Tian, Yuhang Guo, Jian Luan, Bin Wang, and Shuoying Chen. 2023. Cbsimt: Mitigating hallucination in simultaneous machine translation with weighted prefix-to-prefix training. *ArXiv preprint*, abs/2311.03672.

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.

Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020. Monotonic multihead attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. 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 3456–3468, Online. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021a. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021b. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, 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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. 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 1126–1140, Online. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2675–2684, Online. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2022. Probing causes of hallucinations in neural machine translations. *ArXiv preprint*, abs/2206.12529.

Tengfei Yu, Liang Ding, Xuebo Liu, Kehai Chen, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. PromptST: Abstract prompt learning for end-to-end speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10140–10154, Singapore. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2021. Universal simultaneous machine translation with mixture-of-experts wait-k policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7306–7317, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2022a. Information-transport-based policy for simultaneous translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 992–1013, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2022b. Modeling dual read/write paths for simultaneous machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2461–2477, Dublin, Ireland. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2022c. Reducing position bias in simultaneous machine translation with length-aware framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6775–6788, Dublin, Ireland. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang. 2020. Simultaneous translation policies: From fixed to adaptive. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2847–2853, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A Detailed Experimental Settings

On IWSLT'14 De↔En, we train on 160K pairs, develop on 7K held out pairs. All data is tokenized and lower-cased and we segment sequences using byte pair encoding (Sennrich et al., 2016) with 10K merge operations. The resulting vocabularies are of 8.8K and 6.6K types in German and English respectively.

On MuST-C Release V2.0 Zh→En[2], we train on 358,853 pairs, develop on 1,349 pairs. Jieba[3] are employed for Chinese word segmentation. All

[2]https://ict.fbk.eu/must-c-release-v2-0/
[3]https://github.com/fxsjy/jieba

data is tokenized by SentencePiece resulting in 32k word vocabularies in Chinese and English.

Following Elbayad et al. (2020) and Zhang and Feng (2021), We train Transformer Small on IWSLT14 De→En. We train Transformer Base on MuST-C Release V2.0 Zh→En.

B Experimental Results on IWSLT14 En → De Dataset

B.1 Results of Word Frequency Distribution on IWSLT14 De→En Dataset
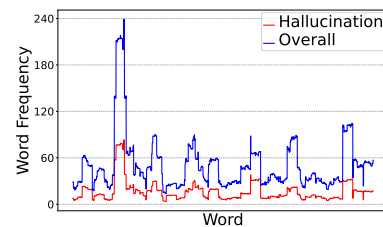
Figure 6: Word frequency of Hallucination and Overall on IWSLT14 De→En valid hypotheses set of wait-1.
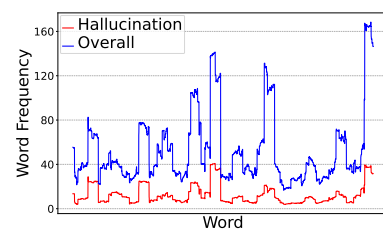

Figure 7: Word frequency of Hallucination and Overall on IWSLT14 De→En valid hypotheses set of wait-3.
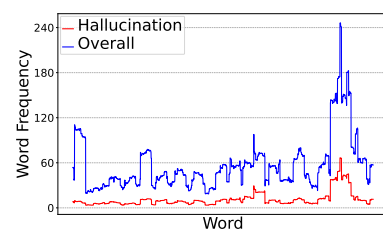

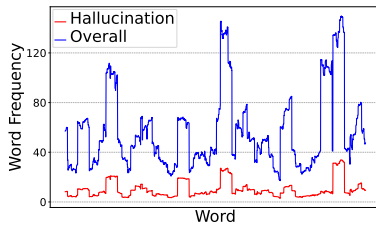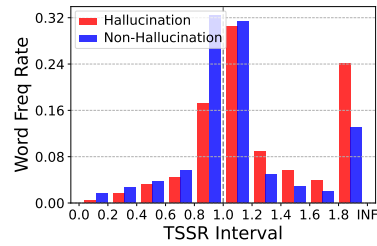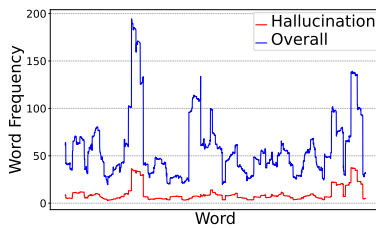Figure 8: Word frequency of Hallucination and Overall on IWSLT14 De→En valid hypotheses set of wait-5.

Figure 9: Word frequency of Hallucination and Overall on IWSLT14 De→En valid hypotheses set of wait-7.

Figure 10: Word frequency of Hallucination and Overall on IWSLT14 De→En valid hypotheses set of wait-9.

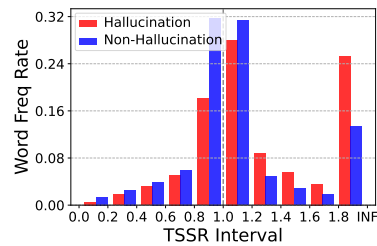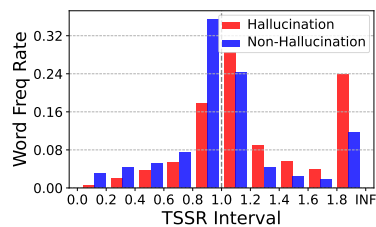## B.2 Results of Word Frequency Rate in TSSR on IWSLT14 De→En Dataset

Figure 11: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-1 model.
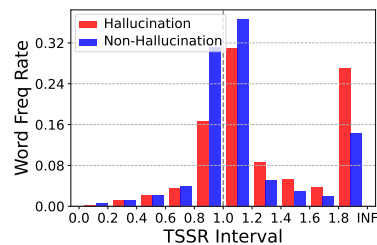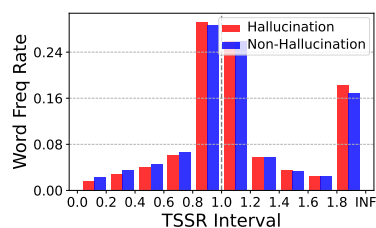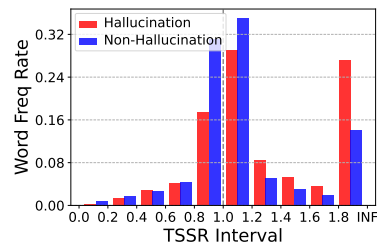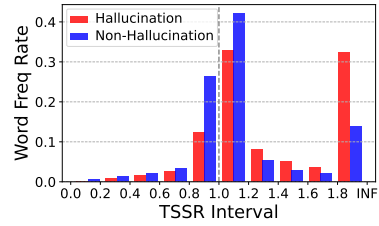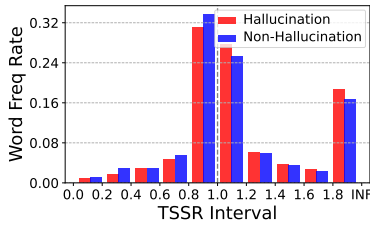

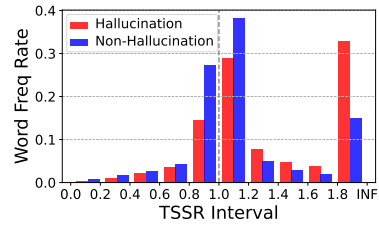Figure 12: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-1 model with WSPAlign Annotation (**?**).

Figure 13: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-3 model.

Figure 14: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-5 model.
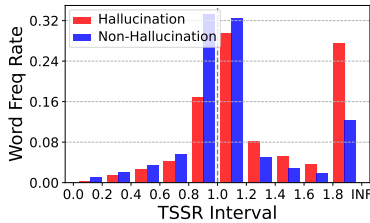

Figure 15: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-7 model.

Figure 16: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for wait-9 model.

738

C Experimental Results on IWSLT14 En→De Dataset
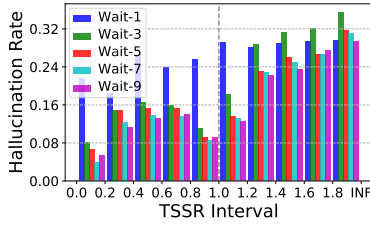
Figure 17: HR on the valid set in different TSSR intervals of wait-k models.

Figure 18: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-1 model.
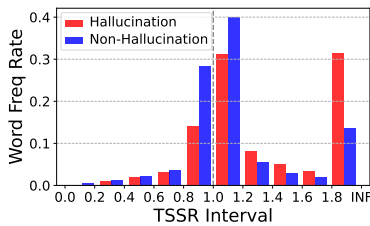

Figure 19: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-3 model.

Figure 20: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-5 model.

Figure 21: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-7 model.

Figure 22: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for wait-9 model.

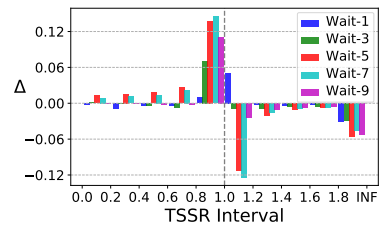|		k=1	k=3	k=5	k=7	k=9
Baselines	BLEU ↑	15.75	22.03	24.99	26.22	26.60
	HR % ↓	27.46	19.73	16.72	16.24	15.93
Scheduled-Sampling	BLEU ↑	16.83	22.78	25.80	26.98	27.41
	HR % ↓	26.19	18.58	15.66	14.96	14.81

Table 5: BLEU scores and HR of wait-k models.

Figure 23: Word Frequency Rate Change (Δ) in different TSSR intervals with scheduled sampling training compared to the Baselines.
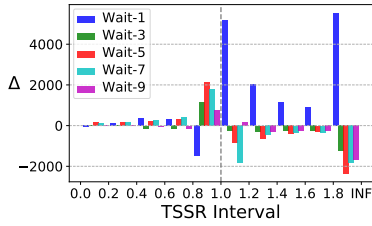
Figure 24: Hallucination Frequency Change (Δ) in different TSSR intervals with scheduled sampling training compared to the Baselines.

D   Experimental Results on MuST-C Zh→En Dataset

k	1	3	5	7	9	∞
HR %	33.96	25.31	23.22	21.84	20.73	19.43

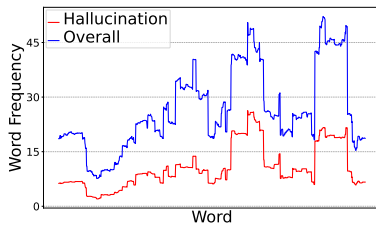Table 6: HR on MuST-C Zh→En valid set of wait-k, where $k = \infty$ means Full-sentence MT.

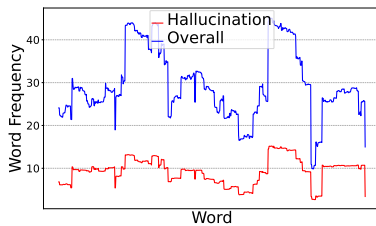Figure 25: Word frequency of Hallucination and Overall on valid hypotheses set of wait-1.

Figure 26: Word frequency of Hallucination and Overall on valid hypotheses set of wait-3.
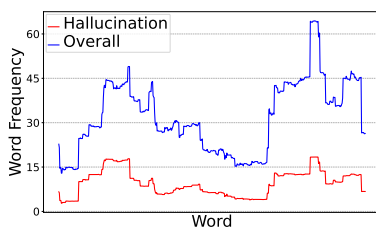

Figure 27: Word frequency of Hallucination and Overall on valid hypotheses set of wait-5.
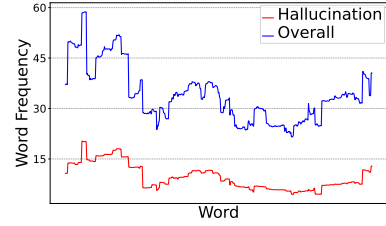

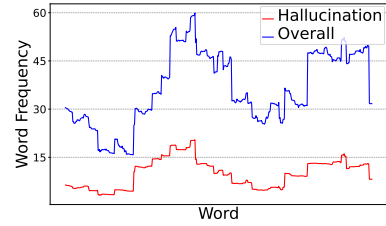Figure 28: Word frequency of Hallucination and Overall on valid hypotheses set of wait-7.

Figure 29: Word frequency of Hallucination and Overall on valid hypotheses set of wait-9.

k	1	3	5	7	9
Hallucination	6.57	6.52	6.35	6.29	6.23
Overall	8.23	8.44	8.49	8.53	8.52

Table 7: Word frequency distribution entropy of Hallucination and Overall on MuST-C Zh→En valid hypotheses set of wait-k.

	Train Ref	Valid Ref	Valid Hypo
Train Ref	1.00	0.25	0.18
Valid Ref	0.25	1.00	0.54
Valid Hypo	0.18	0.54	1.00

Table 8: The correlation between the HR of words on the Valid Hypotheses (Valid Hypo), Valid Reference (Valid Ref) and Train Reference (Train Ref) of $H_{wait-1}(t, a)$.

Wait-k	Valid set				Training subset			
	Uncertainty		Confidence		Uncertainty		Confidence	
	H	NH	H	NH	H	NH	H	NH
k=1	3.23	2.70	0.44	0.54	3.27	2.34	0.44	0.60
k=3	3.00	2.43	0.49	0.58	2.91	2.14	0.50	0.63
k=5	2.67	2.33	0.53	0.60	2.59	2.00	0.55	0.65
k=7	2.64	2.32	0.54	0.60	2.50	2.00	0.56	0.65
k=9	2.60	2.29	0.55	0.60	2.44	2.00	0.57	0.65

Table 9: The Uncertainty and Confidence of Hallucination (**H**) and Non-Hallucination (**NH**) on the valid set and training subset of wait-k models.
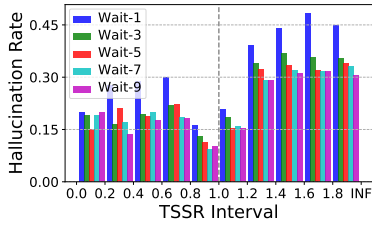
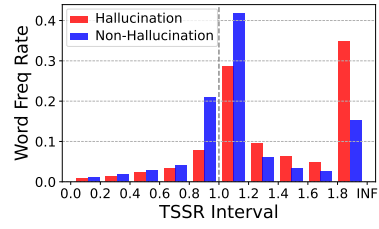Figure 30: HR on the valid set in different TSSR intervals of wait-k models.
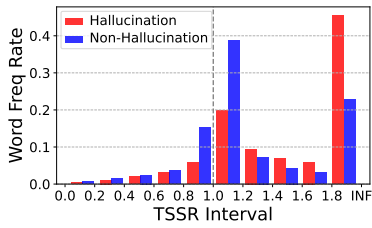

Figure 31: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-1 model.

Figure 32: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-3 model.

Figure 33: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-5 model.

Figure 34: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for the wait-7 model.
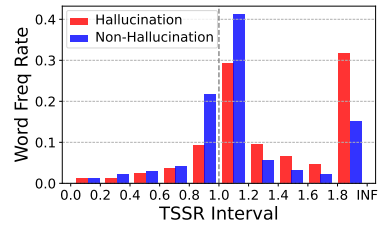

Figure 35: Word Frequency Rate of Hallucination and Non-Hallucination in different TSSR intervals for wait-9 model.
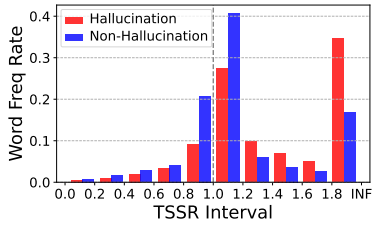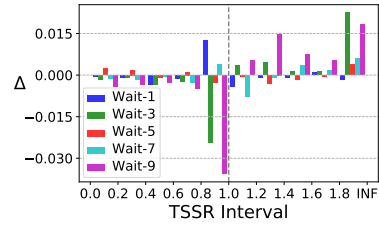

Figure 36: Word Frequency Rate Change (Δ) in different TSSR intervals with scheduled sampling training compared to the Baselines.
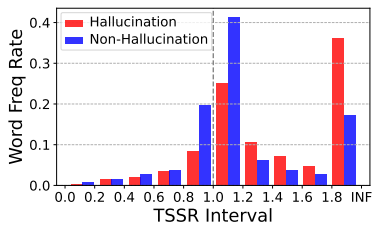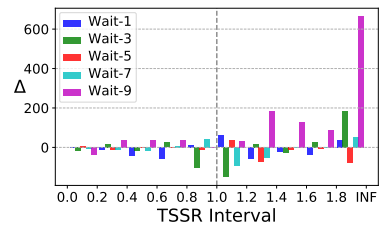

Figure 37: Hallucination Frequency Change (Δ) in different TSSR intervals with scheduled sampling training compared to the Baselines.

741

		k=1	k=3	k=5	k=7	k=9
Baselines	BLEU ↑	12.33	15.39	16.26	16.66	16.66
	HR % ↓	33.96	25.31	23.22	21.84	20.73
Scheduled-Sampling	BLEU ↑	12.42	15.51	16.43	16.61	17.03
	HR % ↓	33.69	25.29	22.68	21.61	23.50

Table 10: BLEU scores and HR of wait-k models.

E Examples of Hallucinations

Source Input	*ér*	*bù shì*	*jiān xīn*	*de*	*sùi yuè*
	而 and	不是 not	艰辛 hard	的	岁月 time
Output		And	not	hard	**work** .

Figure 38: Translation examples of hallucination words as defined in Section 2 under the wait-1 policy. Words highlighted in red indicate hallucinations.

Source Input	*tōng guò*	*hé*	*gōng sī*	*de*	*hé zuò*	
	通过 by	和 with	公司 company	的 's	合作 working	...
Output	**And**	by	**working**	with	**the** company	...

Figure 39: Translation examples of hallucination words as defined in Section 2 under the wait-1 policy. Words highlighted in red indicate hallucinations. when decoding the word "working", the source-side context is "通过 和 公司" and this context lacks the semantic information of "working", as it does not include the aligned word "working" in the current source-side context. Consequently, "working" can be identified as one of the hallucinated words in this output.

Source Input	*xiān rán*	*qí zhōng*	*zhī yī*	*de*	*gǎn shòu*	*bǐ*	*ling*	*yī gè*	*hái*	*chà*	
	显然 Obviously	其中 Among them	之一 One of	的 of	感受 feelings	比 than	另 another	一个 one	还 still	差 worse	.
Output		Obviously	**,**	one	of	them	feels	more	**different**	than	the other .

Figure 40: Translation examples of hallucination words as defined in Section 2 under the wait-1 policy.

Source Input	*nà lǐ*	*de*	*rén*	*xū yào*	*zhè xiē*	
	那里 there	的	人 people	需要 need	这些 these	,
Output	**There** **'s**	**a**	**lot** of	people **out**	there **who**	need it ,

Figure 41: Translation examples of hallucination words as defined in Section 2 under the wait-1 policy.

F Alignment Error Rate of Awesome-Align

Alignment Error Rate	7.30 %
Precision	0.950
Recall	0.885

Table 11: The alignment error rate, precision, and recall of hallucination detection using Awesome-align, with human annotations as the ground truth.

We report the alignment error rate as well as the precision and recall of hallucination detection using Awesome-align. Based on the precision and recall results, we believe that the automatic word alignment is suitable for detecting hallucinated words.