Improving Simultaneous Translation by Incorporating Pseudo-References with Fewer Reorderings

Junkun Chen [1*] **Renjie Zheng** [2*] **Atsuhito Kita** [1†] **Mingbo Ma** [2] **Liang Huang** [1,2]
[1]Oregon State University, Corvallis, OR, USA
[2]Baidu Research, Sunnyvale, CA, USA
chenjun2@oregonstate.edu, renjiezheng@baidu.com

Abstract

Simultaneous translation is vastly different from full-sentence translation, in the sense that it starts translation before the source sentence ends, with only a few words delay. However, due to the lack of large-scale, high-quality simultaneous translation datasets, most such systems are still trained on conventional full-sentence bitexts. This is far from ideal for the simultaneous scenario due to the abundance of unnecessary long-distance reorderings in those bitexts. We propose a novel method that rewrites the target side of existing full-sentence corpora into simultaneous-style translation. Experiments on Zh→En and Ja→En simultaneous translation show substantial improvements (up to +2.7 BLEU) with the addition of these generated pseudo-references.

1 Introduction

Simultaneous translation, which starts translation before the source sentence ends, is substantially more challenging than full-sentence translation due to partial observation of the (incrementally revealed) source sentence. Recently, it has witnessed great progress thanks to fixed-latency policies (such as wait-k) (Ma et al., 2019) and adaptive policies (Gu et al., 2017; Arivazhagan et al., 2019).

However, all state-of-the-art simultaneous translation models are trained on conventional parallel text which involve many unnecessary long-distance reorderings (Birch et al., 2009; Braune et al., 2012); see Fig. 1 for an example. The simultaneous translation models trained using these parallel sentences will learn to either make bold hallucinations (for fixed-latency policies) or introduce long delays (for adaptive ones). Alternatively, one may want to use transcribed corpora from professional simultaneous interpretation (Matsubara et al., 2002; Bendazzoli et al., 2005; Neubig et al., 2018). These data are more monotonic in word-order, but they are all very small in size due to the high cost of data collection (e.g., the NAIST one (Neubig et al., 2018) has only $387k$ target words). More importantly, simultaneous interpreters tend to summarize and inevitably make many mistakes (Shimizu et al., 2014; Xiong et al., 2019; Zheng et al., 2020) due to the high cognitive load and intense time pressure during interpretation (Camayd-Freixas, 2011).

Source Input	*zhōngguó de*	*xībù*	*yǒu*	*hǔnduō*	*gāo*	*shān*	
	中国 的	西部	有	很多	高	山	
	china 's	*west*	*have*	*many*	*big*	*mountain*	
Gold-Ref	**there**	**are**	**many**	**big**	**mountains**	in	**western** china
Pseudo-Refs	*(wait-1)* china 's	**west**	**has**	**many**	**big**	**mtns**	
	(...wait-2...) the	chinese	**west**	**has**	**many**	**big**	**mtns**
	(...wait-3...)	**western**	china	**has**	**many**	**big**	**mtns**
	(...wait-4...)		there	are	**many**	**big**	...

Figure 1: Example of unnecessary reorderings in the bitext which can force the model to anticipate aggressively, along with the ideal pseudo-references with different wait-k policies. Larger k improves fluency but sacrifices latency (pseudo-refs with $k \geq 4$ are identical to the original reference). (mtns: mountains)

How can we combine the merits of both types of data, and obtain a large-scale, more monotonic parallel corpora for simultaneous translation? We propose a simple and effective technique to generate pseudo-references with fewer reorderings; see the "Pseudo-Refs" in Fig. 1. While previous work (He et al., 2015) addresses this problem via language-specific hand-written rules, our technique can be easily adopted to any language pairs without using extra data or expert linguistic knowledge. Training with these generated pseudo references can reduce anticipations during training and result in fewer hallucinations in decoding and lower latency. We make the following contributions:

- We propose a method to generate pseudo-references which are *non-anticipatory* and *semantic preserving*.
- We propose two metrics to quantify the antic-

*Equal contribution. †Currently at Columbia University.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5857–5864
November 7–11, 2021.

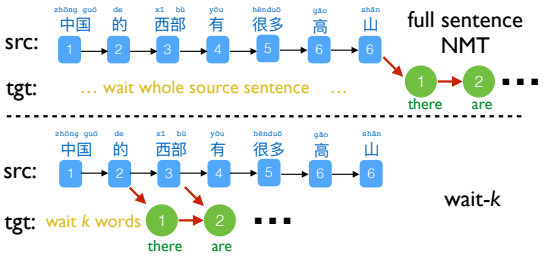

Figure 2: Full-sentence vs. simultaneous (wait-k) MT.

ipation rate in the pseudo-references and the hallucination rate in the hypotheses.

- Our pseudo-references lead to substantial improvements (up to +2.7 BLEU) on Zh→En and Ja→En simultaneous translation.

2 Preliminaries

We briefly review full-sentence neural translation and the wait-k policy in simultaneous translation.

Full-Sentence NMT uses a Seq2seq framework (Fig. 2) where the encoder processes the source sentence $\mathbf{x} = (x_1, x_2, ..., x_m)$ into a sequence of hidden states. A decoder sequentially generates a target sentence $\mathbf{y} = (y_1, y_2, ..., y_n)$ conditioned on those hidden states and previous predictions:

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} p_{\text{full}}(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}^{\text{full}})$$

$$p_{\text{full}}(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = \prod_{t=1}^{|\mathbf{y}|} p(y_t \mid \mathbf{x}, \mathbf{y}_{<t}; \boldsymbol{\theta})$$

The model is trained as follows:

$$\boldsymbol{\theta}^{\text{full}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{(\mathbf{x},\mathbf{y}^*)\in D} p_{\text{full}}(\mathbf{y}^* \mid \mathbf{x}; \boldsymbol{\theta}) \quad (1)$$

Simultaneous Translation translates concurrently with the (growing) source sentence, so Ma et al. (2019) propose the wait-k policy (Fig. 2) following a simple, fixed schedule that commits one target word on receiving each new source word, after an initial wait of k source words. Formally, the prediction of $\mathbf{y}$ for a trained wait-k model is

$$p_{\text{wait-}k}(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = \prod_{t=1}^{|\mathbf{y}|} p(y_t \mid \mathbf{x}_{<t+k}, \mathbf{y}_{<t}; \boldsymbol{\theta}) \quad (2)$$

where the wait-k model is trained as follows

$$\boldsymbol{\theta}^{\text{wait-}k} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{(\mathbf{x},\mathbf{y}^*)\in D} p_{\text{wait-}k}(\mathbf{y}^* \mid \mathbf{x}; \boldsymbol{\theta}).$$

This way, the model learns to implicitly anticipate at testing time, though not always correct (e.g., in Fig. 2, after seeing x_1x_2="中国 的" (China 's), output y_1="there") . The decoder generates the target sentence $\hat{\mathbf{y}}$ with k words behind source sentence $\mathbf{x}$:

$$\hat{y}_t = \operatorname*{argmax}_{y_t} p_{\text{wait-}k}(y_t \mid \mathbf{x}_{<t+k}, \hat{\mathbf{y}}_{<t}; \boldsymbol{\theta}^{\text{wait-}k})$$

3 Pseudo-Reference Generation

Since the wait-k models are trained on conventional full-sentence bitexts, their performance is hurt by unnecessary long-distance reorderings between the source and target sentences. For example, the training sentence pair in Fig. 2, a wait-2 model learns to output y_1="*there*" after observing x_1x_2="中国 的" (*china 's*) which seems to induce a good anticipation ("中国 的..." ↔ "*There ...*"), but it could be a wrong hallucination in many other contexts (e.g., "中国 的 街道 很 挤" ↔ "*Chinese streets are crowded*", not "*There ...*"). Even for adaptive policies (Gu et al., 2017; Arivazhagan et al., 2019; Zheng et al., 2019a), the model only learns a higher latency policy (wait till x_4="有") by training on the example in Fig. 2. As a result, training-time wait-k models tend to do wild hallucinations (Ma et al., 2019).

To solve this problem, we propose to generate pseudo-references which are *non-anticipatory* under a specific simultaneous translation policy by the method introduced in Section 3.1. Meanwhile, we also propose to use BLEU score to filter the generated pseudo-references to guarantee that they are *semantic preserving* in Section 3.2.

3.1 Generating Pseudo-References with Test-time Wait-k

To generate *non-anticipatory* pseudo-references under a wait-k policy, we propose to use the full-sentence NMT model $\boldsymbol{\theta}^{\text{full}}$ (Eq. 1) which is *not* trained to anticipate, but decode with a wait-k policy. This combination is called *test-time wait-k* (Ma et al., 2019), which is unlikely to hallucinate since the full source content is always available during training. Although here the full-sentence model $\boldsymbol{\theta}^{\text{full}}$ only has access to the partially available source words $\boldsymbol{x}_{<t+k}$, it can still enforce fluency because $\hat{y}_t$ relies on the decoded target-side prefix $\hat{\boldsymbol{y}}_{<t}$ (Eq. 2). Formally, the generation of pseudo-references is:

$$\tilde{\mathbf{y}}^* = \operatorname*{argmax}_{\mathbf{y}} p_{\text{wait-}k}(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}^{\text{full}})$$

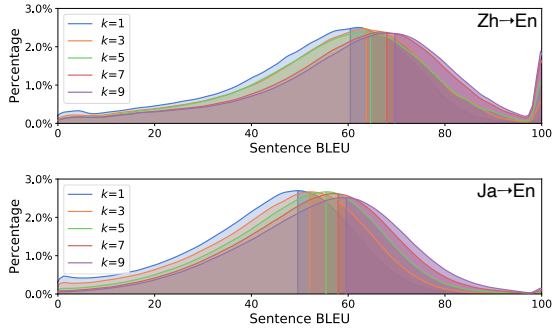

Figure 3: Sentence-level BLEU distributions of Pseudo-Refs using wait-k policies for Zh→En and Ja→En, respectively. The parts to the right of the vertical lines indicate the top 40% references in terms of BLEU in each distribution.

Fig. 1 shows the pseudo-references with different wait-k policies ($k = 1..4$). Note that $k = 1$ or 2 results in non-idiomatic translations, and larger k leads to more fluent pseudo-references, which converge to the original reference with $k \geq 4$. The reason is that in each wait-k policy, each target word $\hat{y}_t$ only rely on observed source words ($\mathbf{x}_{<t+k}$).

To further improve the quality of the pseudo-references generated by test-time wait-k, we propose to select better pseudo-references by using beam search. Beam search usually improves translation quality but its application to simultaneous translation is non-trivial, where output words are committed on the fly (Zheng et al., 2019b). However, for pseudo-reference generation, unlike simultaneous translation decoding, we can simply adopt conventional off-line beam search algorithm since the source sentence is completely known. A larger beam size will generally give better results, but make anticipations more likely to be retained if they are correct and reasonable. To trade-off the expectations of quality and monotonicity, we choose beam size $b = 5$ in this work.

3.2 Translation Quality of Pseudo-References

We can use sentence-level BLEU score to filter out low quality pseudo-references. Fig. 3 shows the sentence level BLEU distributions of the pseudo-references generated with different wait-k policies. As k increases, the translation qualities are better since more source prefixes can be observed during decoding. The obvious peak at the BLEU=100 on Zh→En denotes those pseudo-references which are identical to the original ones. Those original references are probably already non-

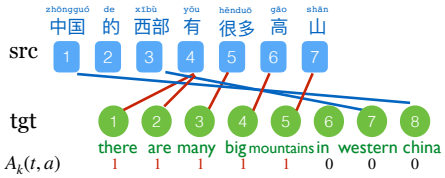

Figure 4: An example word alignment and the wait-1 policy. The red and blue lines indicate the 1-anticipated and non-anticipated alignments, resp. Here AR_1 = 5/8.

hallucinatory or correspond to very short source sentences (e.g. shorter than k). The figure shows that even for wait-1 policy, around 40% pseudo-references can achieve BLEU score above 60.

4 Anticipation & Hallucination Metrics

4.1 Anticipation Rate of (Pseudo-)References

During the training of a simultaneous translation model, an anticipation happens when a target word is generated before the corresponding source word is encoded. To identify the anticipations, we need the word alignment between the parallel sentences.

A word alignment a between a source sentence $\mathbf{x}$ and a target sentence $\mathbf{y}$ is a set of source-target word index pairs (s, t) where the s^{th} source word x_s aligns with the t^{th} target word y_t. In the example in Fig. 4, the word alignment is: $a = \{(1, 8), (3, 7), (4, 1), (4, 2), (5, 3), (6, 4), (7, 5)\}$.

Based on the word alignment a, we propose a new metric called "k-anticipation" to detect the anticipations under wait-k policy. Formally, a target word y_t is k-anticipated ($A_k(t, a) = 1$) if it aligns to at least one source word $\mathbf{x}_s$ where $s \geq t + k$:

$$A_k(t, a) = \mathbb{1}[\{(s, t) \in a \,|\, s \geq t + k\} \neq \varnothing]$$

We further define the k-anticipation rate (AR_k) of an $(\mathbf{x}, \mathbf{y}, a)$ triple under wait-k policy to be:

$$AR_k(\mathbf{x}, \mathbf{y}, a) = \frac{1}{|\mathbf{y}|} \sum_{t=1}^{|\mathbf{y}|} A_k(t, a)$$

4.2 Hallucination Rate of Hypotheses

The goal of reducing the anticipation rate during the training of a simultaneous translation model is to avoid hallucination at testing time. Similar to the anticipation metric introduced in the previous section, we define another metric to quantify the number of hallucinations in decoding. A target word $\hat{y}_t$ is a *hallucination* if it can not be aligned to any source word. Formally, based on word alignment a, whether target word $\hat{y}_t$ is a hallucination

is

$$H(t,a) = \mathbb{1}[\{(s,t) \in a\} = \varnothing]$$

We further define hallucination rate HR as

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

To avoid non-faithful contextual alignments, we use IBM Model 1 (Brown et al., 1993) for HR.

5 Experiments

Dataset and Model We conduct the experiments on two language pairs Zh→En and Ja→En. We use NIST corpus (2M pairs) for Zh→En as training set, and NIST 2006 and NIST 2008 as dev and test set, which contains 616 and 691 sentences with 4 English references respectively. We also collected a set of references annotated by human interpreters with sight-interpreting[1] for the test set. For Ja→En translation, we use ASPEC corpus (3M pairs). Following Morishita et al. (2019), we only use the first 1.5M parallel sentences and discard the rest noisy data. We use the dev and test datasets in ASPEC with 1,790 and 1,812 pairs. We preprocess the data with Mecab (Kudo et al., 2004) as the word segmentation tool and Unidic (Yasuharu et al., 2007) as its dictionary. Consecutive Japanese tokens which only contain Hiragana characters are combined to reduce the redundancy.

The full-sentence model is trained on the original training set. We use *fast_align* (Dyer et al., 2013) as the word aligner (Model 2 for anticipation and Model 1 for hallucination) and train it on the training set. All the datasets are tokenized with BPE (Sennrich et al., 2016). We implement wait-k policies on base Transformer (Vaswani et al., 2017) following Ma et al. (2019) for all experiments.

Results We compare the performance of wait-k models trained on three different settings: (i) original training references only; (ii) original training references with all Pseudo-Refs; (iii) original training references with top 40% Pseudo-Refs in sentence-level BLEU.

Chinese-to-English Table 1 shows the results of Zh→En translation. Compared with using original references only, adding Pseudo-Refs substantially improves the translation quality and reduces hallucination rate. The filtered 40% Pseudo-Refs achieve the best results except $k = 9$. Fig. 7 shows that the generated Pseudo-Refs can significant reduce the k-anticipation rate compared with the original training references, especially for smaller k. As shown in Table 2, if taking the human sight-interpreting result as a single reference, the improvement is more salient than evaluated on the standard 4 references (+7.5% vs. +6.5%), which confirms that our method tend to translate in a "*syntactic linearity*" fashion like human sight and simultaneous interpreters (Ma, 2019).

Fig. 5 shows an example of how the wait-k model is improved by generated Pseudo-Refs. In this example, the original training references actively delay the translation of adverbial clause (time). It makes the model learn to anticipate the subject before its appearance. It is common in the original set. Fig. 6 shows two other examples of generated pseudo references on Ja→En and Zh→En, respectively. The generated pseudo-references are obviously more ideal than the original references. We also show several examples of solving other avoidable anticipations in Figs. A1–A4 in the Appendix.

Japanese-to-English Table 3 shows the results of Ja→En translation task. Japanese-to-English simultaneous translation is a more difficult task due to long distance reorderings (SOV-to-SVO); many Japanese sentences are difficult to be translated into English monotonically. Besides that, the test set has only one single reference and does not cover many possible expressions. Results show that filtered Pseudo-Refs still improve the translation quality (Tab. 3), and reduces anticipation (Fig. 7) and hallucination (Tab. 3).

(4-reference BLEU)		k=1	k=3	k=5	k=7	k=9	Avg.△
Training-Refs (*)	BLEU ↑	29.7	32.1	34.2	35.6	37.6	
	HR% ↓	*8.4*	*7.8*	*6.4*	*6.0*	*5.8*	
*+100% Pseudo-Refs	BLEU ↑	31.8	32.6	35.9	37.9	**39.4**	+1.7 (5.0%)
	HR% ↓	***5.5***	*7.4*	*5.4*	*5.2*	***4.6***	−1.3 (18.9%)
*+Top 40% Pseudo-Refs	BLEU ↑	**32.3**	**34.3**	**36.4**	**38.4**	38.8	+2.2 (6.5%)
	HR% ↓	*5.9*	***5.8***	***5.3***	***5.1***	*5.3*	−1.4 (20.3%)

Table 1: BLEU scores and hallucination rates (HR) of Zh→En wait-k models on the test set against the original 4 references. (Full-sentence BLEU: 39.9).

(single-reference BLEU)	k=1	k=3	k=5	k=7	k=9	Avg.△
Training-Refs (*)	10.9	12.1	13.0	13.7	13.8	
*+Top 40% Pseudo-Refs	**12.6**	**14.2**	**13.9**	**14.2**	**14.1**	+1.1 (7.5%)

Table 2: BLEU scores of Zh→En wait-k models on the test set, taking human sight interpretation as reference.

[1] Sight interpreting refers to (real-time) oral translation of written text. It is considered as a special variant of simultaneous interpretation but with better translation quality.

Training Source Input	*zhōngguó rùshì yǐhòu , zhōng měi liǎng guó jiāng* ...
	中国 入世 以后 , 中 美 两 国 将 ... **(a) Training Example**
	china entry wto after , china USA two country will
(a) Gold Training-Ref	... **the two countires** will ... **after china 's entry into the wto** .
(a') wait-3 Pseudo-Ref	... **after china 's accession to the wto** , **china and the united states** will ...

Dev Source Input	*fēngzhōng hòu , shǒushù qǔdé yuánmǎn chénggōng* 。
	29@@ 5 分钟 后 , 手术 取得 圆满 成功 。 **(b) Dev-set Decoding Results**
	minutes after surgery achieve complete success .
(b) Only Training-Refs	**the two countries** had a complete success in the operation **after 2@@ 95 minutes** .
(b') + top 40% Pseudo-Refs	**2@@ 95 minutes later** , the operation was a complete success .

Figure 5: In the training example in (a), the gold reference anticipates "the two countries", which encourages the wait-k model trained on it to make irrelevant hallucination after any temporal phrase; see the decoding example in (b). Training with the pseudo-reference in (a') fixes this problem, resulting in the correct translation in (b').

Training Source Input	現在 までに 症例 ・ 照 の ２０ ペアが 有効 回答 として 報告 された 。
	Present by case and contrast 20 pairs effective answers as are reported .
Gold Training-Ref	**20 pairs** of case and before contrast **were reported** as a usefulness answers **by the present** .
wait-3 Pseudo-Ref	**to the present** , **20 pairs** of cases and controls **have been reported** as effective answers .

Training Source Input	*jiǎng zuò kāishǐ qián* , lǐ péng fābiǎo jiǎnghuà 。
	讲座 开始 前 , 李 鹏 发表 讲话 。
	lecture begin before , li peng deliver speech .
Gold Training-Ref	**li peng made a speech before the start of the lecture minutes** .
wait-3 Pseudo-Ref	**before the lecture began** , **li peng gave a speech** .

Figure 6: Two examples dealing with adverbial clause delay. The adverbial clauses are at the end of the training references. This introduces anticipation during training and hallucination during decoding.

(single-reference BLEU)		k=3	k=5	k=7	k=9	Avg.△
Training-Refs (*)	BLEU ↑	16.6	19.0	20.8	21.7	
	HR% ↓	*10.8*	*7.3*	*6.5*	*6.2*	
*+100% Pseudo-Refs	BLEU ↑	17.7	18.9	20.8	22.2	+0.3 (1.5%)
	HR% ↓	***6.5***	***6.2***	***5.6***	*5.3*	−1.4 (18.2%)
*+Top 40% Pseudo-Refs	BLEU ↑	**17.9**	**19.2**	**21.5**	**22.5**	+0.6 (3.1%)
	HR% ↓	*8.3*	*7.6*	*6.0*	***5.2***	−0.7 (9.1%)

Table 3: BLEU scores and *HR* of Ja→En wait-k models on the test set. (Full-sentence: 28.4).

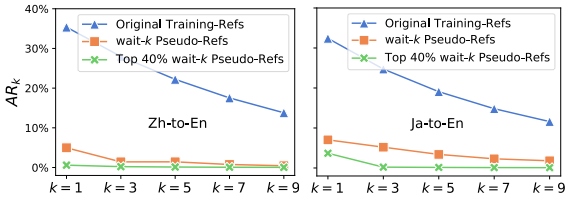

Figure 7: k-Anticipation rates (AR_k) of gold training references and Pseudo-Refs with various k. The top 40% Pseudo-Refs are selected in terms of BLEU.

6 Related Work

In the pre-neural statistical MT era, there exist several efforts using source-side reordering as a preprocessing step for full-sentence translation (Collins et al., 2005; Galley and Manning, 2008; Xu et al., 2009). Unlike this work, they rewrite the source sentences. But in the simultaneous translated scenario, the source input is incrementally revealed and unpredictable. Zheng et al. (2018) propose to improve full sentence translation by generating pseudo-references from multiple gold references, while our work does not require the existence of *multiple* gold references and is designed for simultaneous translation.

This work is closely related to the work of He et al. (2015), which addresses the same problem but only in the special case of Ja→En translation, and uses handwritten language-specific syntactic transformations rules to rewrite the original reference into a more monotonic one. By comparison, our work is much more general in the following aspects: (a) it is not restricted to any language pairs; (b) it does not require language-specific grammar rules or syntactic processing tools; and (c) it can generate pseudo-references with a specific policy according to the requirement of latency.

7 Conclusions

We have proposed a simple but effective method to generate more monotonic pseudo references for simultaneous translation. These pseudo references cause fewer anticipations and can substantially improve simultaneous translation quality.

Acknowledgements

This work is supported in part by NSF IIS-1817231 and IIS-2009071 (L.H.).

References

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323.

Claudio Bendazzoli, Annalisa Sandrelli, et al. 2005. An approach to corpus-based interpreting studies: developing epic (european parliament interpreting corpus). *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation-Saarbrücken*, pages 2–6.

Alexandra Birch, Phil Blunsom, and Miles Osborne. 2009. A quantitative analysis of reordering phenomena. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205. Association for Computational Linguistics.

Fabienne Braune, Anita Gojun, and Alexander Fraser. 2012. Long-distance reordering during search for hierarchical phrase-based smt. In *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, pages 28–30. Citeseer.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Erik Camayd-Freixas. 2011. Cognitive theory of simultaneous interpreting and training. In *Proceedings of the 52nd Conference of the American Translators Association*, volume 13.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 531–540. Association for Computational Linguistics.

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.

Michel Galley and Christopher D Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062.

He He, Alvin Grissom II, John Morgan, Jordan Boyd-Graber, and Hal Daumé III. 2015. Syntax-based rewriting for simultaneous machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 55–64.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 230–237.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. 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.

Xingcheng Ma. 2019. Effect of word order asymmetry on cognitive process of english-chinese sight translation by interpreting trainees: Evidence from eye-tracking.

Shigeki Matsubara, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2002. Bilingual spoken monologue corpus for simultaneous machine interpretation research. In *LREC*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2019. NTT neural machine translation systems at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 99–105, Hong Kong, China. Association for Computational Linguistics.

Graham Neubig, Hiroaki Shimizu, Sakriani Sakti, Satoshi Nakamura, and Tomoki Toda. 2018. The naist simultaneous translation corpus. In *Making Way in Corpus-based Interpreting Studies*, pages 205–215. Springer.

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.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a simultaneous translation corpus for comparative analysis. In *LREC*, pages 670–673. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Dutongchuan: Context-aware translation model for simultaneous interpreting. *arXiv preprint arXiv:1907.12984*.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Josef Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 245–253.

DEN Yasuharu, OGISO Toshinobu, OGURA Hideki, YAMADA Atsushi, MINEMATSU Nobuaki, UCHIMOTO Kiyotaka, and KOISO Hanae. 2007. The development of an electronic dictionary for morphological analysis and its application to japanese corpus linguistics. *Japanese linguistics*, 22:101–123.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019a. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1349–1354.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3188–3197.

Renjie Zheng, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019b. Speculative beam search for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1395–1402.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, Jiahong Yuan, Kenneth Church, and Liang Huang. 2020. Fluent and low-latency simultaneous speech-to-speech translation with self-adaptive training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3928–3937.

A Appendices

Training Source Input	*wǔjiǎodàlóu*	*méiyǒu*	*xuānbù*	*xīn*	*de*	***fāshè***	***rìqí***	。
	五角大楼	没有	宣布	新	的	发射	日期	。
	pentagon	not	announce	new	's	**launch**	**date**	
Gold Training-Ref				no	new	**launch**	**date**	was announced by the pentagon .
wait-3 Pseudo-Ref				the	pentagon has		not	announced a new **launch date** .

Figure A1: The training reference uses passive voice while the source sentence uses active voice. This kind of problem often appears in sentences with "there be" (e.g. Fig. A2). The generated Pseudo-Ref can avoid anticipation by keeping the active voice as the source sentence.

Training Source Input	*liǎng*	*guó*	*jīngmào*	*hézuò*	*cúnzài*	*zhe*	***hěn***	***dà***	***de***	***qiánlì***	。
	两	国	经贸	合作	存在	着	很	大	的	潜力	。
	two	country	economic trade	corperation	exist		**very**	**big**	**'s**	**potential**	.
Gold Training-Ref				there	is	**very great potential**			for	economic and trade cooperation between the two countries .	
wait-3 Pseudo-Ref				the	economic and		trade	cooperation between the			two countries has **great potential** .

Figure A2: A similar example in which the pseudo-reference avoids the anticipation brought by the "there be" phrase in the gold reference.

Training Source Input	*dàn*	*xiéyì*	*hái*	*xūyào*	*dédào*	*sūdān*	*nèigé*	*de*	***pīzhǔn***	。		
	但	协议	还	需要	得到	苏丹	内阁	的	批准	。		
	but	agreement	also	need	get	sudan	cabinet	's	**approval**	.		
Gold Training-Ref				but	the	agreement still		needs	**approval** by the		sud@@ anese cabinet .	
wait-3 Pseudo-Ref				but	the	agreement still		needs to		be **approved** by	the	sud@@ anese cabinet .

Figure A3: The generated Pseudo-Ref avoids anticipation by adding a preposition "to".

Training Source Input	*wǒmen*	*de*	*xīnwén*	*méitǐ*	*nénggòu*	*dédào*	*rénmín*	*de*	*xìnrèn*	,	***gēnběn***	***yuányīn***	*jiù*	*zài*	*zhèlǐ*	.
	我们	的	新闻	媒体	能够	得到	人民	的	信任	,	根本	原因	就	在	这里	。
	we	's	news	media	can	get	people	's	trust	,	**fundamental**	**reason**	that	on	this	.
Gold Training-Ref	this	is	the	**fundamental reason**		why	our	news	media can		be	trust	by	the people		.
wait-3 Pseudo-Ref				our	news	media can		obtain the		trust of		the	people ,		the	**fundamental reason** for this .
wait-5 Pseudo-Ref						our	news	media can		win	the	trust	of	the people, and this is the **fundamental reason** .		

Figure A4: Comparisons of Pseudo-Refs using different wait-k policies. These examples also show the trade-off between latency and fluency of pseudo-references.