

Simultaneous Translation Policies: From Fixed to Adaptive

Baigong Zheng¹ Kaibo Liu¹ Renjie Zheng¹ Mingbo Ma¹
Hairong Liu¹ Liang Huang^{1,2}

¹Baidu Research, Sunnyvale, CA, USA

²Oregon State University, Corvallis, OR, USA

{baigongzheng, kaiboliu, renjiezhang, mingboma}@baidu.com

{liuhairong, lianghuang}@baidu.com

Abstract

Adaptive policies are better than fixed policies for simultaneous translation, since they can flexibly balance the tradeoff between translation quality and latency based on the current context information. But previous methods on obtaining adaptive policies either rely on complicated training process, or underperform simple fixed policies. We design an algorithm to achieve adaptive policies via a simple heuristic composition of a set of fixed policies. Experiments on Chinese→English and German→English show that our adaptive policies can outperform fixed ones by up to 4 BLEU points for the same latency, and more surprisingly, it even surpasses the BLEU score of full-sentence translation in the greedy mode (and very close to beam mode), but with much lower latency.

1 Introduction

Simultaneous translation (ST) aims to provide good translation quality while keeping the latency of translation process as low as possible. This is very important for the scenarios that require simultaneity, such as international summits and negotiations. For this, human interpreters usually start translation before the source sentence ends. However, this makes the translation process much more challenging than the full-sentence translation, because to balance the translation quality and latency, interpreters need to make decisions on when to continue translation and when to stop temporarily to wait for more source side information, which are difficult, especially for syntactically divergent language pairs, such as German and English.

The above decisions can be considered as two actions: READ (wait for a new source word) and WRITE (emit a translated target word) (Gu et al., 2017). Then we only need to decide which action to choose at each step, and the solution can be represented by a *policy*. Earlier works (Yarmohammadi

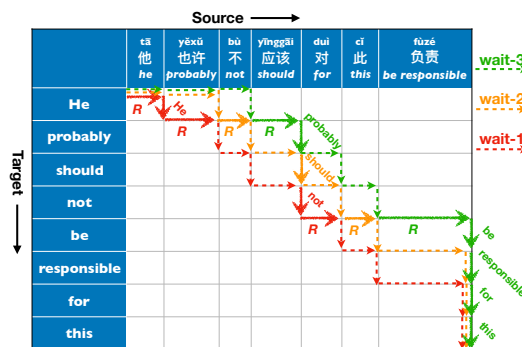


Figure 1: An adaptive policy (in bold arrows) composed of three wait- k policies ($k = 1, 2, 3$).

et al., 2013; Bangalore et al., 2012; Fügen et al., 2007; Sridhar et al., 2013; Jaitly et al., 2016) study policies as a part of speech-to-speech ST system, where the policies usually try to separate the source sentence into several chunks that can be translated safely. Recent works focus on obtaining policies for text-to-text ST, which can be generally divided into two categories: fixed and adaptive. Fixed policies (Ma et al., 2019; Dalvi et al., 2018) usually follow some simple rules to choose actions. For example, the wait- k policy by Ma et al. (2019) first chooses k READ actions, and then chooses WRITE and READ alternatively. This kind of policies do not utilize the context information and can be either too aggressive or too conservative in different cases.

By contrast, adaptive policies try to make decisions on the fly using the currently available information. It is obvious that this kind of policies is more desirable for ST than the fixed ones, and different methods are explored to achieve an adaptive policy. The majority of such methods (Grissom II et al., 2014; Cho and Esipova, 2016; Gu et al., 2017; Alinejad et al., 2018; Zheng et al., 2019a) are based on full-sentence translation models, which may be simple to use but cannot outperform fixed policies applied with “genuinely simultaneous” models trained for ST (Ma et al., 2019). Other meth-

ods (Arivazhagan et al., 2019; Zheng et al., 2019b) try to learn a policy together with the underlying translation model, but they rely on complicated and time-consuming training process.

In this paper, we propose to achieve an adaptive policy via a much simpler heuristic composition of a set of wait- k policies (e.g., $k = 1 \sim 10$). See Fig. 1 for an example. To further improve the translation quality of our method, we apply ensemble of models trained with different wait- k policies. Our experiments on Chinese→English and German→English translation show that our method can achieve up to 4 BLEU points improvement over the wait- k method for same latency. More interestingly, compared with full-sentence translation, our method achieves higher BLEU scores than greedy search but with much lower latency, and is close to the results from beam search.

2 Preliminaries

Full-sentence translation. Neural machine translation (NMT) model usually consists of two components: an encoder, which encodes the source sentence $\mathbf{x} = (x_1, \dots, x_m)$ into a sequence of hidden states, and a decoder, which sequentially predicts target tokens conditioned on those hidden states and previous predictions. The probability of the predicted target sequence $\mathbf{y} = (y_1, \dots, y_n)$ will be

$$p(\mathbf{y} | \mathbf{x}) = \prod_{t=1}^{|\mathbf{y}|} p(y_t | \mathbf{x}, \mathbf{y}_{<t})$$

where $\mathbf{y}_{<t} = (y_1, \dots, y_{t-1})$ denotes the target sequence predicted before step t .

Simultaneous translation. Ma et al. (2019) propose a *prefix-to-prefix* framework to train models to make predictions conditioned on partial source sentences. In this way, the probability of predicted sequence \mathbf{y} becomes

$$p_{\mathbf{g}}(\mathbf{y} | \mathbf{x}) = \prod_{t=1}^{|\mathbf{y}|} p(y_t | \mathbf{x}_{\leq g(t)}, \mathbf{y}_{<t})$$

where $g(t)$ is a monotonic non-decreasing function of t , denoting the number of processed source tokens when predicting y_t . This function $g(t)$ can be used to represent a policy for ST. Ma et al. (2019) introduce a kind of fixed policies, called wait- k policy, that can be defined by the following

$$g_k(t) = \min\{|\mathbf{x}|, t + k - 1\}.$$

Intuitively, this policy first waits k source tokens and then outputs predicted tokens concurrently with the rest of source sentence.

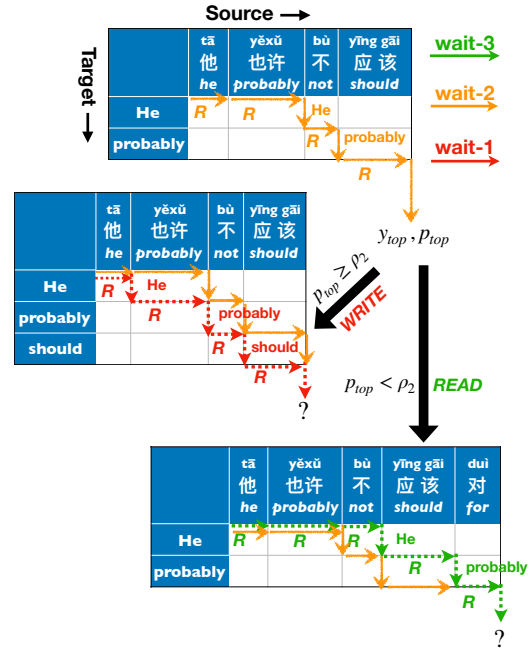


Figure 2: Choose actions based on model confidence. In this example, we will choose an action based on the top probability p_{top} , and apply a new policy (the dotted arrows) after the chosen action.

3 Obtaining an Adaptive Policy

Assume we have a set of wait- k policies and the corresponding models M_k ($k = k_{min} \dots k_{max}$). We can obtain an adaptive policy, whose lag at each step is between k_{min} and k_{max} , meaning that at each step, the target sequence falls behind the source sequence at most k_{max} tokens and at least k_{min} tokens. At each step, there is a wait- k policy synchronizing the adaptive policy, meaning that they have the same lag at that step. Specifically, at any step t , if the lag of the adaptive policy is k' , then we apply the NMT model with the wait- k' policy and force it to predict existing target tokens until step t , when the model will make a new prediction as the output of step t .

However, the above method only shows how to simulate the adaptive policy to make a prediction at one step if we would like to write at that step, but it does not tell us at which steps we should write. We utilize the model confidence to make such a decision. Specifically, we set a probability threshold ρ_k for each wait- k policy. At each step, if the NMT model follows a wait- k' policy, and predicts the most likely token with probability higher than the threshold $\rho_{k'}$, then we consider the model is confident on this prediction, and choose WRITE action; otherwise, we choose READ action. Figure 2 gives an example for this process.

We define the process of applying a wait- k model M_k with a wait- k policy on a given sequence pair (\mathbf{x}, \mathbf{y}) by the following

$$y_{\text{top}}, p_{\text{top}} \leftarrow P_k(M_k, \mathbf{x}, \mathbf{y})$$

which forces model M_k to predict \mathbf{y} , and returns the top token y_{top} at the final step with the corresponding probability p_{top} . The process of reading and returning a new source token is denoted by $READ()$, and expression $\mathbf{x} \circ x$ represents to append an element x to the end of sequence \mathbf{x} . We denote by $\langle s \rangle$ and $\langle /s \rangle$ the start symbol and end symbol of a sequence. Then Algorithm 1 gives the pseudocode of the above method.

Algorithm 1 ST decoding with an adaptive policy

Input: two integers k_{\min} and k_{\max} , a set of NMT models M_k , and a sequence of thresholds ρ_k for $k_{\min} \leq k \leq k_{\max}$.

while $x_{|\mathbf{x}|} \neq \langle /s \rangle$ and $y_{|\mathbf{y}|} \neq \langle /s \rangle$ **do**
 $k \leftarrow |\mathbf{x}| - |\mathbf{y}|$
 $y_{\text{top}}, p_{\text{top}} \leftarrow P_k(M_k, \mathbf{x}, \mathbf{y})$
if $k \geq k_{\max}$ or $(k \geq k_{\min}$ and $p_{\text{top}} \geq \rho_k)$
 $\mathbf{y} \leftarrow \mathbf{y} \circ y_{\text{top}}$ \triangleright Write action
else
 $\mathbf{x} \leftarrow \mathbf{x} \circ READ()$ \triangleright Read action
while $y_{|\mathbf{y}|} \neq \langle /s \rangle$ **do**
 $y_{\text{top}}, p_{\text{top}} \leftarrow P_{k_{\max}}(M_{k_{\max}}, \mathbf{x}, \mathbf{y})$
 $\mathbf{y} \leftarrow \mathbf{y} \circ y_{\text{top}}$ \triangleright Write action
return \mathbf{y}

4 Ensemble of Wait- k Models

Using the corresponding model M_k with each wait- k policies may not give us the best performance. If we have a set of models trained independently with different wait- k policies, then we can apply ensemble of those models (Dietterich, 2000; Hansen and Salamon, 1990) to improve the translation quality, which is also used to improve the translation quality of full-sentence translation (Stahlberg and Byrne, 2017). However, there may be two issues to apply ensemble of all models: (1) the runtime for each prediction could be longer, resulting in higher latency; and (2) the translation accuracy may be worse, for the best model for one policy may give bad performance when doing inference with another policy. To avoid these, we propose to apply ensemble of the top-3 models for each policy. That is, we first generate distribution with the top-3 models independently with the same policy,

and then take the arithmetic average of the three distributions as the final token distribution at that step.

5 Experiments

Datasets and models. We conduct experiments on Chinese→English (ZH→EN) and German→English (DE→EN) translation. For ZH→EN, we use NIST corpus (2M sentence pairs) as training set, NIST 2006 as dev set, and NIST 2008 as test set. For DE→EN, we use WMT15 parallel corpus for training, newstest-2013 for validation and newstest-2015 for testing. All datasets are tokenized and segmented into sub-word units with byte-pair encoding (Sennrich et al., 2016). We take Transformer-base (Vaswani et al., 2017) as our model architecture, and follow Ma et al. (2019) to train our model with wait- k policies for integer $1 \leq k \leq 10$. In the following experiments, we only use catchup (Ma et al., 2019) for DE→EN translation, where we read one additional source token after every 6 predictions. We use BLEU (Papineni et al., 2002) as the translation quality metric, and Average Lagging (AL) (Ma et al., 2019) as the latency metric, which measures the lag behind source in terms of the number of source tokens.

Performance with different policies. We first evaluate the performance of each model with different policies, which helps us to choose models for different policies. Specifically, we apply each model with ten different wait- k policies on dev set to compare the performance. Fig. 3 shows the results of five models. We find the best model for one policy may not be the one trained with that policy. For example, on ZH→EN translation, the best model for wait-1 policy is the one trained with wait-3 policy. Further, there is no one model could achieve the best performance for all policies.

Comparing different methods. We compare our method with others from literature: **wait- k** method (Ma et al., 2019) (train and test models with the same wait- k policy), **test-time wait- k** method (Ma et al., 2019) (apply full-sentence model with wait- k policies), **wait-if-diff** (Cho and Esipova, 2016) (start with s_0 source tokens, choose to read only if top token at t -th step differs from that at $(t - \delta)$ -th step), and **wait-if-worse** (Cho and Esipova, 2016) (start with s_0 source tokens, choose to read only if the top probability at t -th step is smaller than that at $(t - \delta)$ -th step). For **wait-if-diff** we set

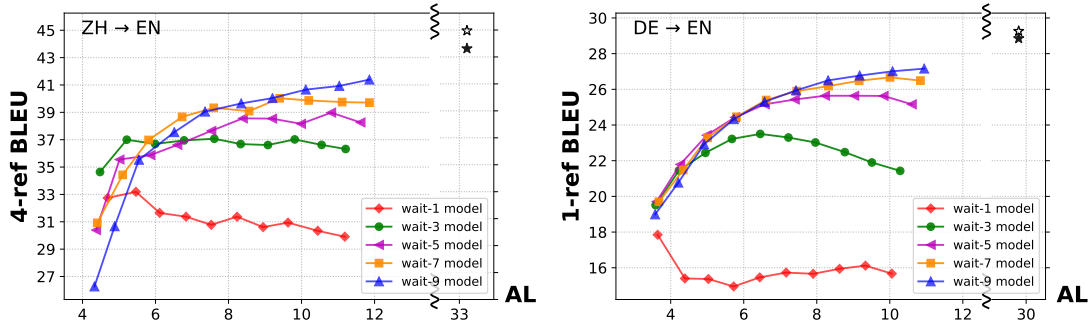


Figure 3: Performance of models with different policies on dev set. Each model is trained with one wait- k policy (i.e. wait- k model) and tested with ten different wait- k' policies for integer $1 \leq k' \leq 10$. Each line corresponds to one model. ★ ☆: full-sentence translation with greedy search and beam search (beam size = 10) respectively.

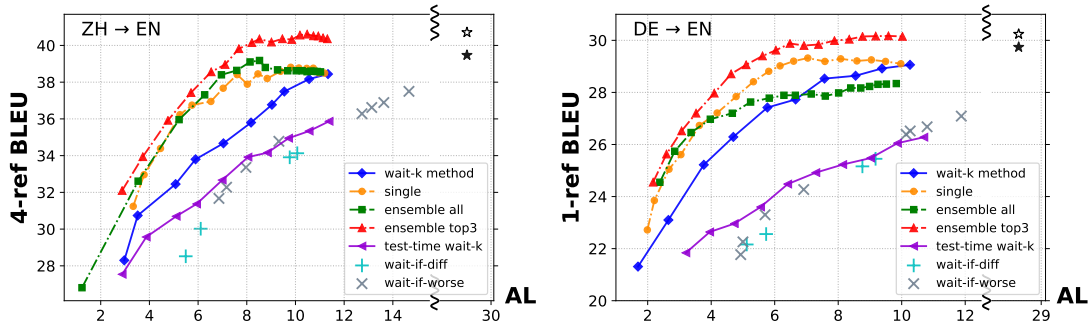


Figure 4: Performance of different methods on test set. Our **single** method achieves better BLEU scores than **wait- k** method with same latency. And our **ensemble top-3** method achieves the highest BLEU scores with same latency, and outperforms full-sentence greedy search with $AL < 9$. ★ ☆: full-sentence translation with greedy search and beam search (beam size = 10) respectively.

$s_0 \in \{4, 6\}$ and $\delta \in \{2, 4\}$; and for **wait-if-worse** we set $s_0 \in \{1, 2, 4, 6\}$ and $\delta \in \{1, 2\}$.

For our method, we test three different cases: (1) **single**, where for each policy we apply the corresponding model that trained with the same policy; (2) **ensemble top-3**, where for each policy we apply the ensemble of 3 models that achieve the highest BLEU scores with that policy on dev set; (3) **ensemble all**, where we apply the ensemble of all 10 models for each policy. For thresholds, we first choose ρ_1 and ρ_{10} , and the other thresholds are computed in the following way: $\rho_i = \rho_1 - d \cdot (i - 1)$ for integer $1 \leq i \leq 10$ and $d = (\rho_1 - \rho_{10})/9$. We test with $\rho_1 \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, $\rho_{10} = 0$ and $\rho_1 = 1$, $\rho_{10} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, totally 18 different settings in our experiments. The reason behind these settings is that we assume our adaptive policy cannot be either too aggressive or too conservative (as mentioned at the beginning of Section 3). The policy is the most aggressive for $k = 1$, so we set ρ_1 as the largest; while for $k = 10$ the policy is the most conservative, so we set ρ_{10} the smallest.

The comparison is provided in Fig. 4 (the corresponding numeric scores are provided in Appendix A). Compared with **wait- k** method, our **single** method achieves improvement of up to 2 BLEU point, and our **ensemble top-3** achieves improvement up to 4 BLEU points. Compared with full-sentence translation, our **ensemble top-3** surprisingly outperforms greedy search with much lower latency ($AL < 9$), and achieves BLEU scores close to that from beam search (see Table 2). We also give one ZH→EN translation example from dev set in Table 1 to compare different methods, showing that our method achieves an adaptive policy with low latency and good translation quality.

Efficiency. To evaluate the efficiency, we present in Table 3 the averaged time needed to predict one token for different methods. These methods are tested on one GeForce GTX TITAN-X GPU for ZH→EN test set. We can see that our **ensemble top-3** method needs about 0.2 seconds to make a prediction on average. However, if the source sentence is revealed in the same speed as general

pinyin input gloss	wǒmen xiàng shòuhàizhě de jiāshǔ biǎoshì zuì chéngzhì de tóngqíng hé āi dào “我们 向 受害者的 家属 表示 最 诚挚 的 同情 和 哀- 悼 . ”	
ensemble top-3 $\rho_1 = 1, \rho_{10} = 0$ (AL=7)	“ we	most sincere express our sympathy and
ensemble top-3 $\rho_1 = 0.4, \rho_{10} = 0$ (AL=2.8)	“ we	express the most sincere sympathy to the families of the victims .
wait-3 (AL=3.72)	“ we have offered our best wishes to the families of the victims	,” he said .
full-sentence translation (AL=16)		“ we express the most sincere sympathy and condolences to the families of the victims . ”

Table 1: One example from ZH→EN dev set. Although **wait-3** method has low latency, it makes anticipations on “offered” and “wishes”, and adds additional words “he said”, which are not accurate translation. Our **ensemble top-3** method could provide better translation with lower latency.

Method	ZH→EN		DE→EN	
	BLEU	AL	BLEU	AL
Full-sentence (greedy)	39.47	29.551	29.74	28.581
Full-sentence (beam)	40.71	29.551	30.24	28.581
Ensemble Top-3	40.15	8.209	30.15	8.766

Table 2: Compare our method with full-sentence translation. Our **ensemble top-3** method could outperform the greedy search and get close to beam search (beam size = 10) with lower latency.

speech, which is about 0.6 seconds per token in Chinese (Zheng et al., 2019c), then our method is still faster than that (which means that it could be used for real-time). Further, we believe the efficiency of our method could be improved with other techniques, such as parallelizing the running of three models in the ensemble, making it less an issue.

Method	Time per Token
Full-sentence	0.0122 s
Wait-3	0.0162 s
Single ($\rho_1 = 0.4, \rho_{10} = 0$)	0.1057 s
Ensemble Top-3 ($\rho_1 = 0.4, \rho_{10} = 0$)	0.2085 s

Table 3: Averaged time needed by different methods to predict one token on ZH→EN test set.

6 Conclusions

We have designed a simple heuristic algorithm to obtain an adaptive policy based on a set of wait- k policies, and applied ensemble in our method to improve the translation quality while maintaining low latency. Experiments show that our method not only outperforms the original wait- k method with relatively large gap, but also surpasses greedy full-sentence translation with much lower latency.

Acknowledgments

We thank the anonymous reviewers for helpful suggestions.

References

- Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar. 2018. [Prediction improves simultaneous neural machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3022–3027.
- 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 Conference of the Association for Computational Linguistics, ACL 2019*, pages 1313–1323.
- Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In *Proc. of NAACL-HLT*.
- Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.
- Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. [Incremental decoding and training methods for simultaneous translation in neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499.
- Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.
- Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine translation*, 21(4):209–252.

- Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don’t until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)*, pages 1342–1352.
- Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O. K. 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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1053–1062.
- Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):993–1001.
- Navdeep Jaitly, David Sussillo, Quoc V Le, Oriol Vinyals, Ilya Sutskever, and Samy Bengio. 2016. An online sequence-to-sequence model using partial conditioning. In *Advances in Neural Information Processing Systems*, pages 5067–5075.
- 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.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, Philadelphia, USA.
- 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.
- Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proc. of NAACL-HLT*, pages 230–238.
- Felix Stahlberg and Bill Byrne. 2017. Unfolding and shrinking neural machine translation ensembles. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1946–1956.
- 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 30*.
- Mahsa Yarmohammadi, Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Baskaran Sankaran. 2013. Incremental segmentation and decoding strategies for simultaneous translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.
- Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019a. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proc. of EMNLP-IJCNLP*, pages 1349–1354.
- Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019b. [Simultaneous translation with flexible policy via restricted imitation learning](#). In *Proc. of ACL*, pages 5816–5822.
- Renjie Zheng, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019c. 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.

A Appendices

We provide the complete results of Figure 4 from Section 5 in the following tables, where AL is Average Lagging. Note that for ZH→EN, we use 4-reference BLEU; while for DE→EN we use single-reference BLEU.

Hyper-parameters		ZH→EN		DE→EN	
		BLEU	AL	BLEU	AL
wait-if-diff	$s_0 = 4, \delta = 2$	28.52	5.493	22.16	5.121
	$s_0 = 6, \delta = 2$	30.02	6.108	22.56	5.731
	$s_0 = 4, \delta = 4$	33.91	9.764	25.16	8.763
	$s_0 = 6, \delta = 4$	34.13	10.075	25.45	9.177
ensemble top-3	$\rho_1 = 0.2, \rho_{10} = 0.0$	32.10	2.880	24.55	2.171
	$\rho_1 = 0.3, \rho_{10} = 0.0$	33.94	3.729	25.63	2.592
	$\rho_1 = 0.4, \rho_{10} = 0.0$	35.92	4.762	26.52	3.068
	$\rho_1 = 0.5, \rho_{10} = 0.0$	37.43	5.710	27.20	3.523
	$\rho_1 = 0.6, \rho_{10} = 0.0$	38.56	6.538	27.97	4.096
	$\rho_1 = 0.7, \rho_{10} = 0.0$	38.96	7.109	28.71	4.628
	$\rho_1 = 0.8, \rho_{10} = 0.0$	39.82	7.675	29.06	5.101
	$\rho_1 = 0.9, \rho_{10} = 0.0$	40.15	8.209	29.40	5.616
	$\rho_1 = 1.0, \rho_{10} = 0.0$	40.35	8.520	29.62	6.038
	$\rho_1 = 1.0, \rho_{10} = 0.1$	40.18	9.013	29.88	6.482
	$\rho_1 = 1.0, \rho_{10} = 0.2$	40.36	9.462	29.80	6.923
	$\rho_1 = 1.0, \rho_{10} = 0.3$	40.32	9.848	29.84	7.379
	$\rho_1 = 1.0, \rho_{10} = 0.4$	40.56	10.185	29.99	7.882
	$\rho_1 = 1.0, \rho_{10} = 0.5$	40.61	10.480	30.04	8.347
	$\rho_1 = 1.0, \rho_{10} = 0.6$	40.52	10.739	30.15	8.766
	$\rho_1 = 1.0, \rho_{10} = 0.7$	40.51	10.939	30.16	9.182
	$\rho_1 = 1.0, \rho_{10} = 0.8$	40.41	11.134	30.17	9.582
	$\rho_1 = 1.0, \rho_{10} = 0.9$	40.36	11.310	30.15	10.023
	ensemble all	$\rho_1 = 0.2, \rho_{10} = 0.0$	26.81	1.231	24.55
$\rho_1 = 0.3, \rho_{10} = 0.0$		32.61	3.536	25.74	2.851
$\rho_1 = 0.4, \rho_{10} = 0.0$		35.96	5.219	26.46	3.367
$\rho_1 = 0.5, \rho_{10} = 0.0$		37.31	6.270	26.97	3.973
$\rho_1 = 0.6, \rho_{10} = 0.0$		38.40	6.959	27.20	4.666
$\rho_1 = 0.7, \rho_{10} = 0.0$		38.64	7.590	27.63	5.241
$\rho_1 = 0.8, \rho_{10} = 0.0$		39.10	8.134	27.78	5.828
$\rho_1 = 0.9, \rho_{10} = 0.0$		39.18	8.523	27.89	6.290
$\rho_1 = 1.0, \rho_{10} = 0.0$		38.80	8.761	27.89	6.650
$\rho_1 = 1.0, \rho_{10} = 0.1$		38.67	9.264	27.94	7.151
$\rho_1 = 1.0, \rho_{10} = 0.2$		38.62	9.682	27.86	7.594
$\rho_1 = 1.0, \rho_{10} = 0.3$		38.62	10.029	27.98	8.014
$\rho_1 = 1.0, \rho_{10} = 0.4$		38.62	10.274	28.17	8.395
$\rho_1 = 1.0, \rho_{10} = 0.5$		38.57	10.477	28.17	8.710
$\rho_1 = 1.0, \rho_{10} = 0.6$		38.60	10.632	28.23	8.989
$\rho_1 = 1.0, \rho_{10} = 0.7$		38.59	10.770	28.31	9.253
$\rho_1 = 1.0, \rho_{10} = 0.8$		38.58	10.890	28.32	9.517
$\rho_1 = 1.0, \rho_{10} = 0.9$		38.56	11.029	28.34	9.830

Table 4: Complete results of **wait-if-diff**, **ensemble top-3** and **ensemble all**.

Hyper-parameters		ZH→EN		DE→EN	
		BLEU	AL	BLEU	AL
wait-if-worse	$s_0 = 1, \delta = 1$	31.67	6.857	21.77	4.930
	$s_0 = 2, \delta = 1$	32.28	7.170	22.26	5.005
	$s_0 = 4, \delta = 1$	33.36	7.964	23.30	5.697
	$s_0 = 6, \delta = 1$	34.78	9.319	24.27	6.914
	$s_0 = 1, \delta = 2$	36.28	12.731	26.52	10.268
	$s_0 = 2, \delta = 2$	36.62	13.133	26.39	10.138
	$s_0 = 4, \delta = 2$	36.89	13.629	26.68	10.806
	$s_0 = 6, \delta = 2$	37.50	14.662	27.09	11.877
single	$\rho_1 = 0.2, \rho_{10} = 0.0$	31.24	3.335	22.72	1.989
	$\rho_1 = 0.3, \rho_{10} = 0.0$	32.96	3.781	23.85	2.211
	$\rho_1 = 0.4, \rho_{10} = 0.0$	34.39	4.455	25.05	2.672
	$\rho_1 = 0.5, \rho_{10} = 0.0$	36.23	5.254	25.61	3.047
	$\rho_1 = 0.6, \rho_{10} = 0.0$	36.75	5.750	26.73	3.627
	$\rho_1 = 0.7, \rho_{10} = 0.0$	36.95	6.526	27.21	4.187
	$\rho_1 = 0.8, \rho_{10} = 0.0$	37.67	7.030	27.84	4.785
	$\rho_1 = 0.9, \rho_{10} = 0.0$	38.41	7.604	28.41	5.330
	$\rho_1 = 1.0, \rho_{10} = 0.0$	37.89	8.021	28.81	5.813
	$\rho_1 = 1.0, \rho_{10} = 0.1$	38.45	8.458	29.02	6.169
	$\rho_1 = 1.0, \rho_{10} = 0.2$	38.20	8.839	29.20	6.596
	$\rho_1 = 1.0, \rho_{10} = 0.3$	38.59	9.386	29.32	7.042
	$\rho_1 = 1.0, \rho_{10} = 0.4$	38.81	9.805	29.19	7.581
	$\rho_1 = 1.0, \rho_{10} = 0.5$	38.77	10.141	29.29	8.079
	$\rho_1 = 1.0, \rho_{10} = 0.6$	38.75	10.463	29.21	8.589
	$\rho_1 = 1.0, \rho_{10} = 0.7$	38.76	10.733	29.25	9.044
	$\rho_1 = 1.0, \rho_{10} = 0.8$	38.51	10.944	29.19	9.491
$\rho_1 = 1.0, \rho_{10} = 0.9$	38.49	11.201	29.10	9.972	
wait-k	$k = 1$	28.30	2.968	21.31	1.695
	$k = 2$	30.74	3.519	23.10	2.652
	$k = 3$	32.45	5.076	25.22	3.768
	$k = 4$	33.80	5.896	26.29	4.697
	$k = 5$	34.67	7.041	27.42	5.771
	$k = 6$	35.80	8.175	27.73	6.658
	$k = 7$	36.77	9.033	28.53	7.569
	$k = 8$	37.49	9.542	28.64	8.548
	$k = 9$	38.17	10.560	28.92	9.379
	$k = 10$	38.44	11.337	29.06	10.261
test-time wait-k	$k = 1$	27.54	2.884	21.84	3.204
	$k = 2$	29.57	3.873	22.64	3.954
	$k = 3$	30.70	5.103	22.96	4.729
	$k = 4$	31.37	5.941	23.60	5.558
	$k = 5$	32.67	6.993	24.48	6.412
	$k = 6$	33.92	8.051	24.92	7.298
	$k = 7$	34.16	8.850	25.23	8.144
	$k = 8$	34.95	9.720	25.48	9.025
	$k = 9$	35.34	10.566	26.05	9.867
	$k = 10$	35.87	11.383	26.28	10.699

Table 5: Complete results of **wait-if-worse**, **single**, **wait-k** and **test-time wait-k**.