Appendix: FlowSeq

A Flow Layers

ActNorm

$$\mathbf{z}'_t = \mathbf{s} \odot \mathbf{z}_t + \mathbf{b}.$$

Log-determinant:

$$T \cdot \mathrm{sum}(\log |\mathbf{s}|)$$

Invertible Linear

$$\mathbf{z}'_t = \mathbf{z}_t \mathbf{W},$$

Log-determinant:

$$T \cdot h \cdot \log |\det(\mathbf{W})|$$

where h is the number of heads.

Affine Coupling

$$\begin{aligned} \mathbf{z}_a, \mathbf{z}_b &= \mathrm{split}(\mathbf{z}) \\ \mathbf{z}'_a &= \mathbf{z}_a \\ \mathbf{z}'_b &= \mathrm{s}(\mathbf{z}_a, \mathbf{x}) \odot \mathbf{z}_b + \mathrm{b}(\mathbf{z}_a, \mathbf{x}) \\ \mathbf{z}' &= \mathrm{concat}(\mathbf{z}'_a, \mathbf{z}'_b), \end{aligned}$$

Log-determinant:

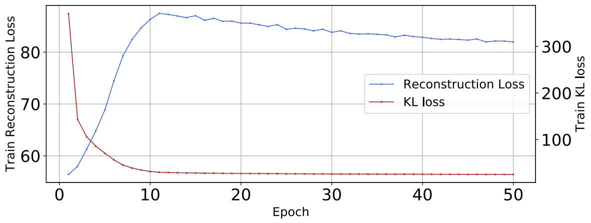
$$\mathrm{sum}(\log |\mathbf{s}|)$$

B Model Details

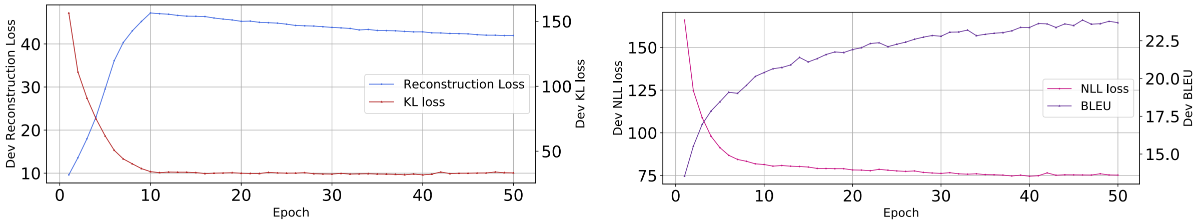
Model	**Dimensions (Model/Hidden)**	**#Params**
Transformer-base	512/2048	65M
Transformer-large	2014/4096	218M
FlowSeq-base	256/512	73M
FlowSeq=large	512/2014	258M

Table 3: Comparison of model size in our experiments.

C Analysis of training dynamics



(a) training loss



(b) dev loss

(c) dev BLEU

Figure 7: Training dynamics.

In Fig. 7, we plot the train and dev loss together with dev BLEU scores for the first 50 epochs. We can see that the reconstruction loss is increasing at the initial stage of training, then start to decrease when training with full KL loss. In addition, we observed that FlowSeq does not suffer the KL collapse problem (Bowman et al., 2015; Ma et al., 2019). This is because the decoder of FlowSeq is non-autogressive, with latent variable $\mathbf{z}$ as the only input.

D Analysis of Translation Results

Source	Grundnahrungsmittel gibt es schlielich berall und jeder Supermarkt hat mittlerweile Sojamilch und andere Produkte.
Ground Truth	There are basic foodstuffs available everywhere , and every supermarket now has soya milk and other products.
Sample 1	After all, there are basic foods everywhere and every supermarket now has soya amch and other products.
Sample 2	After all, the food are available everywhere everywhere and every supermarket has soya milk and other products.
Sample 3	After all, basic foods exist everywhere and every supermarket has now had soy milk and other products.
Source	Es kann nicht erklären, weshalb die National Security Agency Daten ber das Privatleben von Amerikanern sammelt und warum Whistleblower bestraft werden, die staatliches Fehlverhalten offenlegen.
Ground Truth	And, most recently, it cannot excuse the failure to design a simple website more than three years since the Affordable Care Act was signed into law.
Sample 1	And recently, it cannot apologise for the inability to design a simple website in the more than three years since the adoption of Affordable Care Act.
Sample 2	And recently, it cannot excuse the inability to design a simple website in more than three years since the adoption of Affordable Care Act.
Sample 3	Recently, it cannot excuse the inability to design a simple website in more than three years since the Affordable Care Act has passed.
Source	Doch wenn ich mir die oben genannten Beispiele ansehe, dann scheinen sie weitgehend von der Regierung selbst gewählt zu sein.
Ground Truth	Yet, of all of the examples that I have listed above, they largely seem to be of the administration’s own choosing.
Sample 1	However, when I look at the above mentioned examples, they seem to be largely elected by the government itself.
Sample 2	But if I look at the above mentioned examples, they seem to have been largely elected by the government itself.
Sample 3	But when I look at the above examples, they seem to be largely chosen by the government itself.
Source	Damit wollte sie auf die Gefahr von noch gröeren Ruinen auf der Schweizer Wiese hinweisen - sollte das Riesenprojekt eines Tages scheitern.
Ground Truth	In so doing they wanted to point out the danger of even bigger ruins on the Schweizer Wiese - should the huge project one day fail.
Sample 1	In so doing, it wanted to highlight the risk of even greater ruins on the Swiss meadow - the giant project should fail one day.
Sample 2	In so doing, it wanted to highlight the risk of even greater ruins on the Swiss meadow - if the giant project fail one day.
Sample 3	In doing so, it wanted point out the risk of even greater ruins on the Swiss meadow - the giant project would fail one day.

Table 4: Examples of translation outputs from FlowSeq-base with sampling hyperparameters $l = 3, r = 10, \tau = 0.4$ on WMT14-DEEN.

In Tab. 4, we present randomly picked translation outputs from the test set of WMT14-DEEN. For each German input sentence, we pick three hypotheses from 30 samples. We have the following observations: First, in most cases, it can accurately express the meaning of the source sentence, sometimes in a different way from the reference sentence, which cannot be precisely reflected by the BLEU score. Second, by controlling the sampling hyper-parameters such as the length candidates l, the sampling temperature τ and the number of samples r under each length, FlowSeq is able to generate diverse translations expressing the same meaning. Third, repetition and broken translations also exist in some cases due to the lack of language model dependencies in the decoder.

E Results of Translation Diversity

Table 5 shows the detailed results of translation deversity.

Models	τ	**Pairwise BLEU**	**LOO BLEU**
Human	–	35.48	69.07
Sampling	–	24.10	37.80
Beam Search	–	73.00	69.90
Hard-MoE	–	50.02	63.80
FlowSeq l=1, r=10	0.1	79.39	61.61
	0.2	72.12	61.05
	0.3	67.85	60.79
	0.4	64.75	60.07
	0.5	61.12	59.54
	1.0	43.53	52.86
FlowSeq l=2, r=5	0.1	70.32	60.54
	0.2	66.45	60.21
	0.3	63.72	59.81
	0.4	61.29	59.47
	0.5	58.49	58.80
	1.0	42.93	52.58
FlowSeq l=3, r=4	0.1	62.21	58.70
	0.2	59.74	58.59
	0.3	57.57	57.96
	0.4	55.66	57.45
	0.5	53.49	56.93
	1.0	39.75	50.94

Table 5: Translation diversity results of FlowSeq-large model on WMT14 EN-DE with knowledge distillation.