A Categorization of DA Methods

By characterizing the second term in Eq. 3, recent DA methods can be conceptually summarized in Table 4. In RAML, given an instance from the empirical data distribution $\hat{p}$, the augmented tgt is sampled independently without considering the equivariance of the src $\tilde{\mathrm{x}} = \mathrm{x}$ as well as its linguistic smoothness. The guarantee of not incurring much noise is the concentration property of $p_{\tilde{\mathrm{x}}|\mathrm{x}}$ around x. SO modifies RAML to consider augmenting the src as well. This kind of noise injection paradigm has been previously studied by Wager et al. (2013); Wang et al. (2013); Xie et al. (2017) as feature/data noising, and can be seen as a kind of regularization technique as Dropout (Srivastava et al., 2014), also pointed out in Wang et al. (2018). So we call them data noising (dn) based DA methods.

For the ST, TA, BT and DA4Low methods, there exists a translation model trained on different views of the bilingual corpus (different translation direction or decoding order as in ST, BT and TA) or different parameterization (the SMT-like model in DA4Low), which can guarantee the equivariance and smoothness properties of $(\tilde{\mathrm{x}}, \tilde{\mathrm{y}})$. That is, the original model is trained on the outputs from another translation model, which matches the knowledge distillation paradigm (Hinton et al., 2015; Kim and Rush, 2016; Furlanello et al., 2018). So we call them knowledge distillation (kn) based DA methods.

Besides the above DA methods, most other DA methods can be a variant of one of them. For kn bsaed methods, reconstruction (Cheng et al., 2016) is like BT when used on monilingual data with instance reweighting; and when the augmentation process is seen as a stage in iterative co-training between the target model and the augmentation model, dual learning (Xia et al., 2016), and joint training (Zhang et al., 2018) can be unified.

B Translation Tasks and Training Details

Table 5 shows the statistics of the three standard benchmarks we rely on, with the IWSLT corpus for training both translation directions so as to obtain four translation tasks. We choose corpora with different sizes: 0.22M, 1M and 4.5M. All the corpora are publicly available from their websites. [4] [5] [6] To note that we choose the datum2017 corpus as a subset of the Zh⇒En corpus for constructing the medium sized translation task. All the data is pre-processed with Byte Pair Encoding (Sennrich et al., 2016b) by jointly learning the source and target vocabulary. [7]

Table 6 shows the hyper-parameters of training on each translation tasks. In Section 2.1, we have identified two factors to control the effect of learning from $\mathcal{T}$ and $\mathcal{A}$. We conduct experiment among DA methods where the two factors are the same or at least similar. According to the categorization discussed in Appx. A. The dn based methods are doing online augmentation so the interpolation coefficient α equals to the probability an instance is to be augmented. This quantity is derived to be around 0.6. So for ST, TA, BT, we use beam search to augment every instance in the train with the top one decoded instance, so that the α is around 0.5 which is comparable to dn based methods. We do not consider DA4Low in our experiment since the α is around 0.05 which are far from 0.6.

C Kendall's Coefficient of Concordance

The Kendall rank correlation coefficient is computed through the following formula:

$$W = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_i^n X_i)^2}{n}}{\frac{1}{12} \cdot k^2 \cdot (n^3 - n)}, \tag{9}$$

where k is the number of rankings and n the number of objects. In our setting, k is 4 corresponding to the four translation tasks and n is 6 corresponding to the 5 DA methods plus the baseline.

D Measure Binned Avg. Freq. Statistics

This appendix section demonstrates the measure (input sensitivity or prediction margin) binned statistics of our two measures on the other translation tasks in Figure 3 and 4 respectively. The x-axis is one of the measures and the y-axis is the average token frequency within that bin. For the both the measure, we can also see similar trends that both of the measures are improved largely for low frequency tokens instead of high-frequency tokens. Actually, we find that most of the high frequency tokens sacrifice their sensitivity or margin as a trade-off with low frequency tokens.

[4] https://wit3.fbk.eu/mt.php?release=2017-01-trnted

[5] http://nlp.nju.edu.cn/cwmt-wmt/

[6] http://www.statmt.org/wmt19/translation-task.html

[7] https://github.com/rsennrich/subword-nmt

Method		Generative story of $\tilde{\mathrm{x}}, \tilde{\mathrm{y}}$ with q_{AUG}		src/tgt	dn/kn	o/f
RAML	Norouzi et al. (2016)	$\tilde{\mathrm{x}} = \mathrm{x}$;	$\tilde{\mathrm{y}} \sim p_{\tilde{\mathrm{y}}\|\mathrm{y}}$	tgt	dn	o
SO	Wang et al. (2018)	$\tilde{\mathrm{x}} \sim p_{\tilde{\mathrm{x}}\|\mathrm{x}}$;	$\tilde{\mathrm{y}} \sim p_{\tilde{\mathrm{y}}\|\mathrm{y}}$	both	dn	o
ST	Zhang and Zong (2016)	$\tilde{\mathrm{x}} = \mathrm{x}$;	$\tilde{\mathrm{y}} \sim p_{\theta',l2r}(\cdot\|\tilde{\mathrm{x}})$	tgt	kn	f
TA	Zhang et al. (2019)	$\tilde{\mathrm{x}} = \mathrm{x}$;	$\tilde{\mathrm{y}} \sim p_{\theta',r2l}(\cdot\|\tilde{\mathrm{x}})$	tgt	kn	f
BT	Sennrich et al. (2016a)	$\tilde{\mathrm{y}} = \mathrm{y}$;	$\tilde{\mathrm{x}} \sim p_{\theta',l2r}(\cdot\|\tilde{\mathrm{y}})$	src	kn	f
DA4Low	Fadaee et al. (2017)	$\tilde{\mathrm{y}} \sim p_{\gamma,lm}(\cdot\|\mathrm{y})$;	$\tilde{\mathrm{x}} \sim p_{\theta',smt}(\cdot\|\tilde{\mathrm{y}})$	both	kn	f

Table 4: Categorization of various DA methods according to the augmentation distribution. Note that each DA method is given a name abbreviation (RAML: Reward Augmented MLE, SO: Switchout, ST: Self-Training, TA: Target-side Agreement regularization, BT: Back-Translation). The **generative story** column describes specific choices of q_{AUG} and the generation of an augmented instance $(\tilde{\mathrm{x}}, \tilde{\mathrm{y}})$. Here the sampling process $\mathrm{x}, \mathrm{y} \sim \hat{p}$ is omitted since it exists in every DA method. $p_{\theta'}$ is another NMT model trained with different translation direction (conditioned on src/tgt) or different decoding order (l2r or r2l). p_{γ} is might be a well trained bidirectional language model from which we can sample and replace sub-spans in y. The **src/tgt** column shows the side of language a DA method augments. The **dn/kn** column classifies a DA method into data noising based or knowledge distillation based. The **o/f** column classifies a DA method into online or offline augmentation.

Tasks	train	dev	test	BPE merge no.	src vocab	tgt vocab
Fr⇔En	218878	9948	9487	10000	11981	9840
Zh⇒En	998668	3002	3981	60000	46953	37071
En⇒De	4542403	3000	3003	40000	39996	39996

Table 5: Corpus statistics of the four translation tasks.

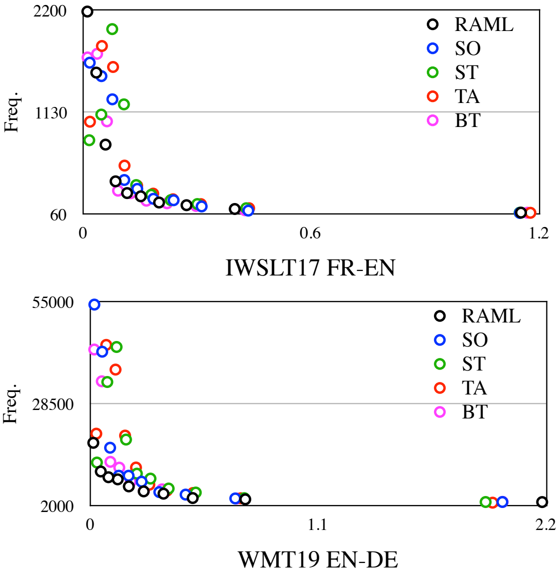

Figure 3: Δ sensitivity binned average token frequency statistics on IWSLT17 Fr⇒En (0.22M) and WMT19 En⇒De (4.5M).

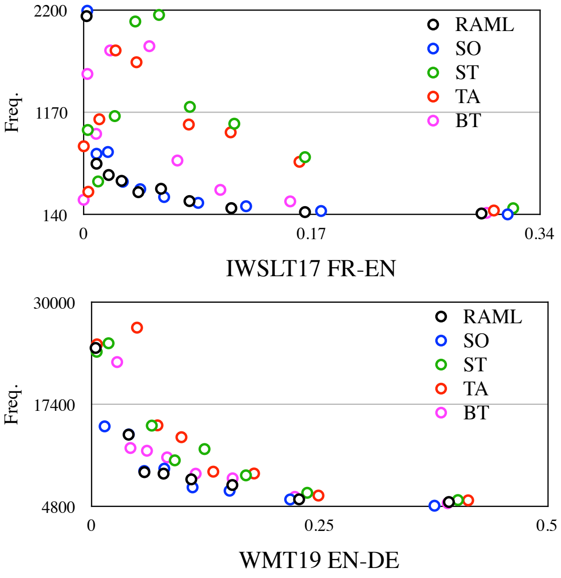

Figure 4: Δ margin binned average token frequency statistics on IWSLT17 Fr⇒En (0.22M) and WMT19 En⇒De (4.5M).

Tasks	n_{layers}	n_{head}	d_{model}	d_{inner}	sched.	n_{warmup}	n_{epoch}	init. lr
Fr⇒En	2	4	256	512	Switchout	-	80	0.001
En⇒Fr	2	4	256	512	Switchout	-	80	0.001
Zh⇒En	6	8	512	2048	inverse_sqrt	4000	30	0.0007
En⇒De	6	8	512	2048	inverse_sqrt	4000	35	0.0007

Table 6: Hyper-parameters for the model and the training.