A Appendices

A.1 Pivot Translation

Recent state-of-the-art NMT models are heavily dependent on a large number of bilingual language resources. Large-sized bilingual text datasets are usually readily available between common and other languages; however, for language pairs that are used less frequently, few or no bilingual resources may be available.

Pivot translation was proposed to overcome the resource limitations for certain language pairs. Recent state-of-the-art NMT models heavily depend on a large number of bilingual language resources. Large-sized bilingual text data sets are usually readily available between the lingua francas and other languages. However, for less-frequently used language pairs, only a limited amount or even none of the bilingual resources are available. Pivot translation was proposed to overcome resource limitations for certain language pairs due to the lack of bilingual language resources. Instead of a direct translation between two languages for which few or no bilingual resources are available, the pivot translation approach makes use of a third language (namely the pivot language). This third language is more appropriate because of the availability of more bilingual corpora and/or its relatedness to either the source or the target language.

Pivot translation has long been studied in statistical machine translation (Wu and Wang, 2007; Utiyama and Isahara, 2007; Paul et al., 2009), supervised NMT (Cheng et al., 2017; Liu et al., 2018; Kim et al., 2019), and UNMT (Leng et al., 2019) as a means of improving the performance of low/zero-resource translations.

Formally, for the translation from language $\mathcal{S}$ to $\mathcal{T}$, the chosen pivot language is denoted as $\mathcal{P}$. The translation schema can be described as follows:

$$\mathcal{S} \rightarrow \mathcal{P}_1 \rightarrow ... \rightarrow \mathcal{P}_K \rightarrow \mathcal{T}, \tag{10}$$

where K is the number of pivot languages.

Recently, the development of UNMT seems to have lessened the importance of pivot translation. UNMT no longer requires bilingual parallel data between two languages, so the low/zero-resource translation problem for less-common language pairs is partially solved; however, the performance of UNMT between some distant languages in different language groups or families is still not promising, which leads researchers to reconsider pivot translation based on UNMT.

	en-fr-ro			
	en→ro	*ro→en*	*fr→ro*	*ro→fr*
UNMT	34.45	32.42	25.26	27.99
MUNMT	34.44	32.60	25.31	27.91
+ RAT-D	34.71	33.01	25.42	28.04
+ RAT-ID	35.83	33.52	25.66	28.25
MUNMT + RNMT	36.39	33.85	25.53	28.57
+ RAT-D	36.43	34.55	25.50	28.59
+ RAT-ID	36.65	34.07	25.78	28.63

Table 5: Comparison of the proposed different RAT implementations.

A.2 RAT-D and RAT-ID

In this paper, the RAT method is proposed to seek the consistency of the outputs of the two translation directions, $\mathcal{S} \rightarrow \mathcal{T}$ and $\mathcal{R} \rightarrow \mathcal{T}$, when their input is parallel. In addition to the implementation described in this paper, the output distribution of $\mathcal{S} \rightarrow \mathcal{T}$ and $\mathcal{R} \rightarrow \mathcal{T}$ can also be directly computed as the agreement loss between $\mathcal{S} \rightarrow \mathcal{T}$ and $\mathcal{R} \rightarrow \mathcal{T}$. For convenience, we call this implementation RAT-D, and we call the implementation described in this paper RAT-ID.

As the two translations $\tilde{\boldsymbol{t}}_{\mathcal{S}}$ and $\tilde{\boldsymbol{t}}_{\mathcal{R}}$ from the parallel sentence pair $\langle \boldsymbol{s}, \boldsymbol{r} \rangle$ should be the same, it is clear that their probability distributions $\tilde{\boldsymbol{d}}_{\mathcal{S}} = \mathbb{P}(\cdot|\boldsymbol{s}; \theta_{\mathcal{S}\rightarrow\mathcal{T}})$ and $\tilde{\boldsymbol{d}}_{\mathcal{R}} = \mathbb{P}(\cdot|\boldsymbol{r}; \theta_{\mathcal{R}\rightarrow\mathcal{T}})$ should ideally be consistent. We would like to minimize the distance of $\tilde{\boldsymbol{d}}_{\mathcal{S}}$ and $\tilde{\boldsymbol{d}}_{\mathcal{R}}$ so that the agreement is learned by the model. The Jensen–Shannon divergence (JSD) (Fuglede and Topsoe, 2004) is then used to compute the difference in the two distributions as the loss for RAT-D training. This is a symmetrized and smoothed version of the Kullback–Leibler divergence (KLD):

$$\begin{aligned}\mathcal{L}_{\textbf{RAT-D}}(\mathcal{S}, \mathcal{T}, \mathcal{R}) = \textbf{JSD}(\tilde{\boldsymbol{d}}_{\mathcal{S}} || \tilde{\boldsymbol{d}}_{\mathcal{R}}) = \\ \frac{1}{2}(\textbf{KLD}(\tilde{\boldsymbol{d}}_{\mathcal{S}} || \boldsymbol{M}) + \textbf{KLD}(\tilde{\boldsymbol{d}}_{\mathcal{R}} || \boldsymbol{M})),\end{aligned} \tag{11}$$

where $\boldsymbol{M} = \frac{1}{2}(\tilde{\boldsymbol{d}}_{\mathcal{S}} + \tilde{\boldsymbol{d}}_{\mathcal{R}})$, and the KLD of distribution Q from P is defined as:

$$\textbf{KLD}(P||Q) = \sum_i P_i \log(\frac{P_i}{Q_i}). \tag{12}$$

Autoregressive NMT models generate translations from left-to-right and stop when an **EOS** token is generated or the generation exceeds the maximum length. This leads to some length inconsistency between the two translation sequences and makes the distributions incompatible for Equation 11. Therefore, in the

training phase, we force the translation model to generate a sequence of length J, which is determined as follows:

$$J = \frac{1}{2}((\alpha J_{\mathcal{S}} + \beta) + (\alpha J_{\mathcal{R}} + \beta)), \tag{13}$$

where $J_{\mathcal{S}}$ and $J_{\mathcal{R}}$ are the lengths of the source language and reference language sentences, respectively; we set $\alpha = 1.3$ and $\beta = 5$ following previous work (Conneau and Lample, 2019).

Differences RAT-D and RAT-ID are the same in principle; both attempt to move two independent output distributions closer to the (weighted) average distribution through the agreement mechanism. The difference is that RAT-D is applied to the two output distributions directly; the two models are required to generate fixed-length distributions before calculating the loss, and there is no interaction between the models in the generation process. The latter point causes an error propagation problem, whereby different errors made in the two translation processes make the context in each translation increasingly different, resulting in two distributions that differ significantly. RAT-ID addresses this issue by obtaining an agreed-upon output prediction at each step, which ensures the context remains consistent in the two model generation processes.

It is shown that the effect of RAT-D is not significant compared to that of RAT-ID, which validates our belief that error propagation caused inconsistent context in the generation we analyzed.