Modeling Bilingual Conversational Characteristics for Neural Chat Translation

Yunlong Liang[1]*, Fandong Meng[2], Yufeng Chen[1], Jinan Xu[1]† and Jie Zhou[2]
[1]Beijing Key Lab of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{yunlongliang,chenyf,jaxu}@bjtu.edu.cn
{fandongmeng,withtomzhou}@tencent.com

Abstract

Neural chat translation aims to translate bilingual conversational text, which has a broad application in international exchanges and cooperation. Despite the impressive performance of sentence-level and context-aware Neural Machine Translation (NMT), there still remain challenges to translate bilingual conversational text due to its inherent characteristics such as role preference, dialogue coherence, and translation consistency. In this paper, we aim to promote the translation quality of conversational text by modeling the above properties. Specifically, we design three latent variational modules to learn the distributions of bilingual conversational characteristics. Through sampling from these learned distributions, the latent variables, tailored for role preference, dialogue coherence, and translation consistency, are incorporated into the NMT model for better translation. We evaluate our approach on the benchmark dataset BConTrasT (English⇔German) and a self-collected bilingual dialogue corpus, named BMELD (English⇔Chinese). Extensive experiments show that our approach notably boosts the performance over strong baselines by a large margin and significantly surpasses some state-of-the-art context-aware NMT models in terms of BLEU and TER. Additionally, we make the BMELD dataset publicly available for the research community.[1]

1 Introduction

A conversation may involve participants that speak in different languages (*e.g.*, one speaking in English and another in Chinese). Fig. 1 shows an example, where the English role R_1 and the Chinese role R_2 are talking about the "*boat*". The goal of chat translation is to translate bilingual conversational text, *i.e.*, converting one participant's language (*e.g.*, English) to another's (*e.g.*, Chinese) and vice versa (Farajian et al., 2020). It enables multiple speakers to communicate with each other in their native languages, which has a wide application in industry-level services.

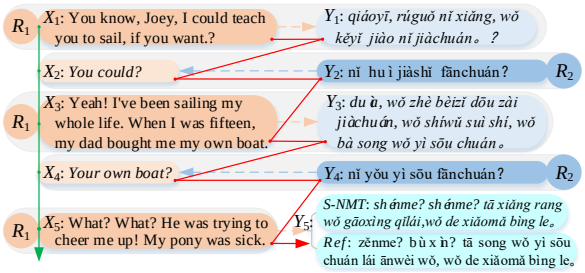

Figure 1: An ongoing bilingual conversation example (English⇔Chinese), where the Chinese utterances are presented in pinyin style. R_i: Role i. The dashed arrows mark the translation direction. The green and red arrows represent the monolingual and bilingual conversation flow, respectively. Although the translation of Y_5 produced by the "*S-NMT*" (a context-free sentence-level NMT system) is reasonable at the sentence level, the coherence of the entire dialogue translation is poor.

Although sentence-level Neural Machine Translation (NMT) (Sutskever et al., 2014; Vaswani et al., 2017; Meng and Zhang, 2019; Hassan et al., 2018; Yan et al., 2020; Zhang et al., 2019) has achieved promising progress, it still faces challenges in accurately translating conversational text due to abandoning the dialogue history, which leads to role-irrelevant, incoherent and inconsistent translations (Mirkin et al., 2015; Wang et al., 2017a; Läubli et al., 2018; Toral et al., 2018). Further, context-aware NMT (Tiedemann and Scherrer, 2017; Voita et al., 2018, 2019a,b; Wang et al., 2019; Maruf and Haffari, 2018; Maruf et al., 2019; Ma et al., 2020) can be directly applied to chat translation through incorporating the dialogue history but cannot obtain satisfactory results in this sce-

*Work was done when Yunlong Liang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
†Jinan Xu is the corresponding author.
[1]Code and data are publicly available at: https://github.com/XL2248/CPCC

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5711–5724
August 1–6, 2021.

nario (Moghe et al., 2020). One important reason is the lack of explicitly modeling the inherent bilingual conversational characteristics, *e.g.*, role preference, dialogue coherence, and translation consistency, as pointed out by Farajian et al. (2020).

For a conversation, its dialogue history contains rich role preference information such as emotion, style, and humor, which is beneficial to role-relevant utterance generation (Wu et al., 2020). As shown in Fig. 1, the utterances X_1, X_3 and X_5 from role R_1 always have strong emotions (*i.e.*, *joy*) because of his/her preference, and preserving the same preference information across languages can help raise emotional resonance and mutual understanding (Moghe et al., 2020). Meanwhile, there exists semantic coherence in the conversation, as the solid green arrow in Fig. 1, where the utterance X_5 naturally and semantically connects with the dialogue history ($X_{1\sim4}$) on the topic "*boat*". In addition, the bilingual conversation exhibits translation consistency, where the correct lexical choice to translate the current utterance might have appeared in preceding turns. For instance, the word "*sail*" in X_1 is translated into "*jiàchuán*", and thus the word "*sailing*" in X_3 should be mapped into "*jiàchuán*" rather than other words (*e.g.*, "*hángxíng*"[2]) to maintain translation consistency. On the contrary, if we ignore these characteristics, translations might be role-irrelevant, incoherent, inconsistent, and detrimental to further communication like the translation produced by the "*S-NMT*" in Fig. 1. Although the translation is acceptable at the sentence level, it is abrupt at the bilingual conversation level.

Apparently, how to effectively exploit these bilingual conversational characteristics is one of the core issues in chat translation. And it is challenging to implicitly capture these properties by just incorporating the complex dialogue history into encoders due to lacking the relevant information guidance (Farajian et al., 2020). On the other hand, the Conditional Variational Auto-Encoder (CVAE) (Sohn et al., 2015) has shown its superiority in learning distributions of data properties, which is often utilized to model the diversity (Zhao et al., 2017), coherence (Wang and Wan, 2019) and users' personalities (Bak and Oh, 2019), etc. In spite of its success, adapting it to chat translation is non-trivial, especially involving multiple tailored latent variables.

[2]The words "*jiàchuán*" and "*hángxíng*" express similar meaning.

Therefore, in this paper, we propose a model, named CPCC, to capture role preference, dialogue coherence, and translation consistency with latent variables learned by the CVAE for neural chat translation. CPCC contains three specific latent variational modules to learn the distributions of role preference, dialogue coherence, and translation consistency, respectively. Specifically, we firstly use one role-tailored latent variable, sampled from the learned distribution conditioned only on the utterances from this role, to preserve preference. Then, we utilize another latent variable, generated by the distribution conditioned on source-language dialogue history, to maintain coherence. Finally, we leverage the last latent variable, generated by the distribution conditioned on paired bilingual conversational utterances, to keep translation consistency. As a result, these tailored latent variables allow our CPCC to produce role-specific, coherent, and consistent translations, and hence make the bilingual conversation go fluently.

We conduct experiments on WMT20 Chat Translation dataset: BConTrasT (En⇔De[3]) (Farajian et al., 2020) and a self-collected dialogue corpus: BMELD (En⇔Ch). Results demonstrate that our model achieves consistent improvements in four directions in terms of BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), showing its effectiveness and generalizability. Human evaluation further suggests that our model effectively alleviates the issue of role-irrelevant, incoherent and inconsistent translations compared to other methods. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to incorporate the role preference, dialogue coherence, and translation consistency into neural chat translation.
- We are the first to build a bridge between the dialogue and machine translation via conditional variational auto-encoder, which effectively models three inherent characteristics in bilingual conversation for neural chat translation.
- Our approach gains consistent and significant performance over the standard context-aware baseline and remarkably outperforms some state-of-the-art context-aware NMT models.
- We contribute a new bilingual dialogue corpus (BMELD, En⇔Ch) with manual translations and our codes to the research community.

[3]English⇔German: En⇔De. English⇔Chinese: En⇔Ch.

2 Background

2.1 Sentence-Level NMT

Given an input sentence $X=\{x_i\}_{i=1}^{M}$ with M tokens, the model is asked to produce its translation $Y=\{y_i\}_{i=1}^{N}$ with N tokens. The conditional distribution of the NMT is:

$$p_\theta(Y|X) = \prod_{t=1}^{N} p_\theta(y_t|X, y_{1:t-1}),$$

where θ are model parameters and $y_{1:t-1}$ is the partial translation.

2.2 Context-Aware NMT

Given a source context $D_X=\{X_i\}_{i=1}^{J}$ and a target context $D_Y=\{Y_i\}_{i=1}^{J}$ with J aligned sentence pairs (X_i, Y_i), the context-aware NMT (Ma et al., 2020) is formalized as:

$$p_\theta(D_Y|D_X) = \prod_{i=1}^{J} p_\theta(Y_i|X_i, X_{<i}, Y_{<i}),$$

where $X_{<i}$ and $Y_{<i}$ are the preceding context.

2.3 Variational NMT

The variational NMT model (Zhang et al., 2016) is the combination of CVAE (Sohn et al., 2015) and NMT. It introduces a random latent variable $\mathbf{z}$ into the NMT conditional distribution:

$$p_\theta(Y|X) = \int_{\mathbf{z}} p_\theta(Y|X, \mathbf{z}) \cdot p_\theta(\mathbf{z}|X) d\mathbf{z}. \quad (1)$$

Given a source sentence X, a latent variable $\mathbf{z}$ is firstly sampled by the prior network from the encoder, and then target sentence is generated by the decoder: $Y \sim p_\theta(Y|X, \mathbf{z})$, where $\mathbf{z} \sim p_\theta(\mathbf{z}|X)$.

As it is hard to marginalize Eq. 1, the CVAE training objective is a variational lower bound of the conditional log-likelihood:

$$\begin{aligned}\mathcal{L}(\theta, \phi; X, Y) &= -\mathrm{KL}(q_\phi(\mathbf{z}|X, Y) \| p_\theta(\mathbf{z}|X)) \\ &\quad + \mathbb{E}_{q_\phi(\mathbf{z}|X,Y)}[\log p_\theta(Y|\mathbf{z}, X)] \\ &\leq \log p(Y|X),\end{aligned}$$

where ϕ are parameters of the posterior network and $\mathrm{KL}(\cdot)$ indicates Kullback–Leibler divergence between two distributions produced by prior networks and posterior networks (Sohn et al., 2015; Kingma and Welling, 2013).

3 Chat NMT

We aim to learn a model that can capture inherent characteristics in the bilingual dialogue history for producing high-quality translations, *i.e.*, using the context for better translations (Farajian et al., 2020). Following (Maruf et al., 2018), we define paired bilingual utterances (X_i, Y_i) as a turn in Fig. 2, where we will translate the current utterance X_{2k+1} at the $(2k+1)$-th turn. Here, we denote the utterance X_{2k+1} as X_u and its translation Y_{2k+1} as Y_u for simplicity, where $X_u=\{x_i\}_{i=1}^{m}$ with m tokens and $Y_u=\{y_i\}_{i=1}^{n}$ with n tokens. Formally, the conditional distribution for the current utterance is

$$p_\theta(Y_u|X_u, C) = \prod_{t=1}^{n} p_\theta(y_t|X_u, y_{1:t-1}, C),$$

where C is the bilingual dialogue history.

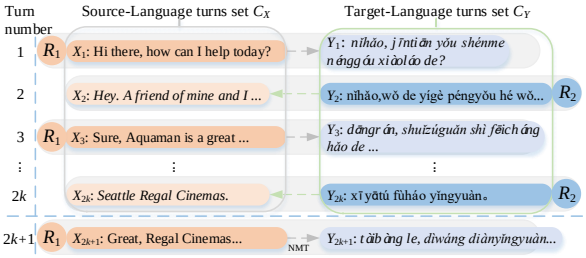

Figure 2: A dialogue example (En⇔Ch) when translating the utterance X_{2k+1} where $k \in [0, \frac{|T|-1}{2}]$ and T is the total number of turns (assumed to be odd here).

Before we dig into the details of how to utilize C, we define three types of context in C (as shown in Fig. 2): (1) the set of previous role-specific source-language turns, denoted as $C_X^{role}=\{X_1, X_3, X_5, ..., X_{2k+1}\}$[4] where $k \in [0, \frac{|T|-3}{2}]$ and T is the total number of turns; (2) the set of previous source-language turns, denoted as $C_X=\{X_1, X_2, X_3, ..., X_{2k}\}$; and (3) the set of previous target-language turns, denoted as $C_Y=\{Y_1, Y_2, Y_3, ..., Y_{2k}\}$.

4 Our Methodology

Fig. 3 demonstrates an overview of our model, consisting of five components: *input representation*, *encoder*, *latent variational modules*, *decoder*, and *training objectives*. Specifically, we aim to model both dialogue and translation simultaneously. Therefore, for the *input representation* (§ 4.1), we incorporate dialogue-level embeddings, *i.e.*, role and dialogue turn embeddings, into the *encoder* (§ 4.2). Then, we introduce three specific *latent variational modules* (§ 4.3) to learn the distributions for varied inherent bilingual characteristics. Finally, we elaborate on how to incorporate the three tailored latent variables sampled from

[4] $C_Y^{role}=\{Y_2, Y_4, Y_6, ..., Y_{2k}\}$ is also role-specific utterances of the interlocutor, which is used to model the interlocutor's consistency in the reverse translation direction. Here, we take one translation direction (*i.e.*, En⇒Ch) as an example.

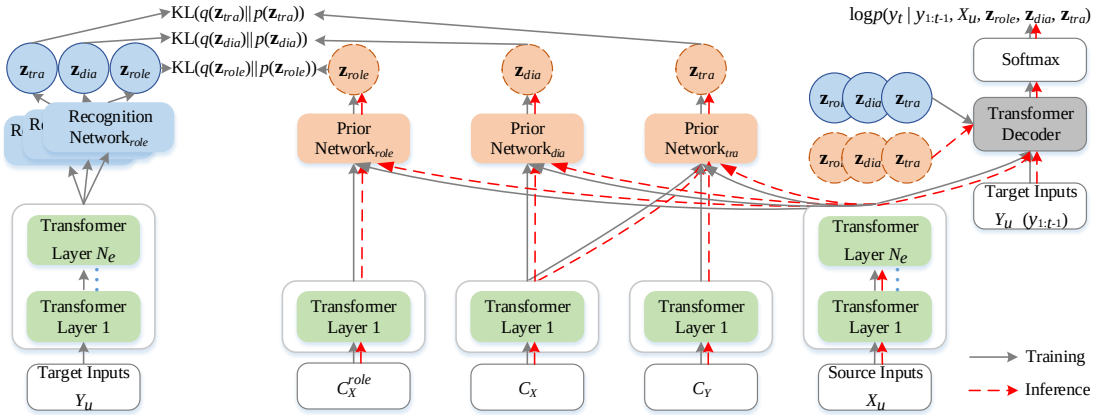

Figure 3: Overview of our CPCC. The latent variables $\mathbf{z}_{role}$, $\mathbf{z}_{dia}$, and $\mathbf{z}_{tra}$ are tailored for maintaining the role preference, dialogue coherence, and translation consistency, respectively. The solid grey lines indicate training process responsible for generating $\{\mathbf{z}_{role}, \mathbf{z}_{dia}, \mathbf{z}_{tra}\}$ from the corresponding posterior distribution predicted by recognition networks. The dashed red lines indicate inference process for generating $\{\mathbf{z}_{role}, \mathbf{z}_{dia}, \mathbf{z}_{tra}\}$ from the corresponding prior distributions predicted by prior networks. The first Transformer layer is shared with all inputs.

the distributions into the *decoder* (§ 4.4) and our two-stage *training objectives* (§ 4.5).

4.1 Input Representation

The CPCC contains three types of inputs: source input X_u, target input Y_u, and context inputs $\{C_X^{role}, C_X, C_Y\}$. Apart from the conventional word embeddings $\mathbf{WE}$ and position embeddings $\mathbf{PE}$ (Vaswani et al., 2017), we also introduce role embeddings $\mathbf{RE}$ and dialogue turn embeddings $\mathbf{TE}$ to identify different utterances. Specifically, for X_u, we firstly project it into these embeddings. Then, we perform a sum operation to unify them into a single input for each token x_i:

$$\mathbf{h}_i^0 = \mathbf{WE}(x_i) + \mathbf{PE}(x_i) + \mathbf{RE}(x_i) + \mathbf{TE}(x_i), \quad (2)$$

where $1 \le i \le m$ and $\mathbf{WE} \in \mathbb{R}^{|V|\times d}$, $\mathbf{RE} \in \mathbb{R}^{|R|\times d}$ and $\mathbf{SE} \in \mathbb{R}^{|T|\times d}$. $|V|$, $|R|$, $|T|$, and d denote the size of shared vocabulary, number of roles, max turns of dialogue, and hidden size, respectively. $\mathbf{h}^0 \in \mathbb{R}^{m\times d}$, similarly for Y_u. For each of $\{C_X^{role}, C_X, C_Y\}$, we add '[cls]' tag at the head of it and use '[sep]' tag to separate its utterances (Devlin et al., 2019), and then get its embeddings via Eq. 2.

4.2 Encoder

The Transformer encoder consists of N_e stacked layers and each layer includes two sub-layers:[5] a multi-head self-attention (SelfAtt) sub-layer and a position-wise feed-forward network (FFN) sub-layer (Vaswani et al., 2017):

$$\mathbf{s}_e^\ell = \text{SelfAtt}(\mathbf{h}_e^{\ell-1}) + \mathbf{h}_e^{\ell-1}, \ \mathbf{h}_e^{\ell-1} \in \mathbb{R}^{m\times d},$$
$$\mathbf{h}_e^\ell = \text{FFN}(\mathbf{s}_e^\ell) + \mathbf{s}_e^\ell, \ \{\mathbf{h}_e^\ell, \mathbf{s}_e^\ell\} \in \mathbb{R}^{m\times d},$$

[5]We omit the layer normalization for simplicity, and you may refer to (Vaswani et al., 2017) for more details.

where $\mathbf{h}_e^\ell$ denotes the state of the ℓ-th encoder layer and $\mathbf{h}_e^0$ denotes the initialized feature $\mathbf{h}^0$.

We prepare the representations of X_u and $\{C_X^{role}, C_X, C_Y\}$ for training prior and recognition networks. For X_u, we apply *mean-pooling* with mask operation over the output $\mathbf{h}_e^{N_e,X}$ of the N_e-th encoder layer, *i.e.*, $\mathbf{h}_X = \frac{1}{m}\sum_{i=1}^m(\mathbf{M}_i^X \mathbf{h}_{e,i}^{N_e,X})$, $\mathbf{h}_X \in \mathbb{R}^d$, where $\mathbf{M}^X \in \mathbb{R}^m$ denotes the mask matrix, whose value is either 1 or 0 indicating whether the token is padded (Zhang et al., 2016). For C_X^{role}, as shown in Fig. 3, we follow (Ma et al., 2020) and share the first encoder layer to obtain the context representation. Here, we take the hidden state of '[cls]' as its representation, denoted as $\mathbf{h}_{role}^{ctx} \in \mathbb{R}^d$. Similarly, we obtain representations of C_X and C_Y, denoted as $\mathbf{h}_X^{ctx} \in \mathbb{R}^d$ and $\mathbf{h}_Y^{ctx} \in \mathbb{R}^d$, respectively.

For training recognition networks, we obtain the representation of Y_u as $\mathbf{h}_Y = \frac{1}{n}\sum_{i=1}^n(\mathbf{M}_i^Y \mathbf{h}_{e,i}^{N_e,Y})$, $\mathbf{h}_Y \in \mathbb{R}^d$, where $\mathbf{M}^Y \in \mathbb{R}^n$, similar to $\mathbf{M}^X$.

4.3 Latent Variational Modules

We design three tailored latent variational modules to learn the distributions of inherent bilingual conversational characteristics, *i.e.*, role preference, dialogue coherence, and translation consistency.

Role Preference. To preserve the role preference when translating the role's current utterance, we only encode the previous utterances of this role and produce a role-tailored latent variable $\mathbf{z}_{role} \in \mathbb{R}^{d_z}$, where d_z is the latent size. Inspired by (Wang and Wan, 2019), we use isotropic Gaussian distribution as the prior distribution of $\mathbf{z}_{role}$: $p_\theta(\mathbf{z}_{role}|X_u, C_X^{role}) \sim \mathcal{N}(\boldsymbol{\mu}_{role}, \boldsymbol{\sigma}_{role}^2\mathbf{I})$, where $\mathbf{I}$

denotes the identity matrix and we have

$$\boldsymbol{\mu}_{role} = \mathrm{MLP}_{\theta}^{role}(\mathbf{h}_X; \mathbf{h}_{role}^{ctx}),$$
$$\boldsymbol{\sigma}_{role} = \mathrm{Softplus}(\mathrm{MLP}_{\theta}^{role}(\mathbf{h}_X; \mathbf{h}_{role}^{ctx})),$$

where $\mathrm{MLP}(\cdot)$ and $\mathrm{Softplus}(\cdot)$ are multi-layer perceptron and approximation of ReLU function, respectively. $(\cdot;\cdot)$ indicates concatenation operation.

At training, the posterior distribution conditions on both role-specific utterances and the current translation, which contain rich role preference information. Therefore, the prior network can learn a role-tailored distribution by approaching the posterior network via KL divergence (Sohn et al., 2015): $q_{\phi}(\mathbf{z}_{role}|X_u, C_X^{role}, Y_u) \sim \mathcal{N}(\boldsymbol{\mu}'_{role}, \boldsymbol{\sigma}'^2_{role}\mathbf{I})$ and $\{\boldsymbol{\mu}'_{role}, \boldsymbol{\sigma}'_{role}\}$ are calculated as:

$$\boldsymbol{\mu}'_{role} = \mathrm{MLP}_{\phi}^{role}(\mathbf{h}_X; \mathbf{h}_{role}^{ctx}; \mathbf{h}_Y),$$
$$\boldsymbol{\sigma}'_{role} = \mathrm{Softplus}(\mathrm{MLP}_{\phi}^{role}(\mathbf{h}_X; \mathbf{h}_{role}^{ctx}; \mathbf{h}_Y)).$$

Dialogue Coherence. To maintain the coherence in chat translation, we encode the entire source-language utterances and then generate a latent variable $\mathbf{z}_{dia} \in \mathbb{R}^{d_z}$. Similar to $\mathbf{z}_{role}$, we define its prior distribution as: $p_{\theta}(\mathbf{z}_{dia}|X_u, C_X) \sim \mathcal{N}(\boldsymbol{\mu}_{dia}, \boldsymbol{\sigma}^2_{dia}\mathbf{I})$ and $\{\boldsymbol{\mu}_{dia}, \boldsymbol{\sigma}_{dia}\}$ are calculated as:

$$\boldsymbol{\mu}_{dia} = \mathrm{MLP}_{\theta}^{dia}(\mathbf{h}_X; \mathbf{h}_X^{ctx}),$$
$$\boldsymbol{\sigma}_{dia} = \mathrm{Softplus}(\mathrm{MLP}_{\theta}^{dia}(\mathbf{h}_X; \mathbf{h}_X^{ctx})).$$

At training, the posterior distribution conditions on both the entire source-language utterances and the translation that provide a dialogue-level coherence clue, and is responsible for guiding the learning of the prior distribution. Specifically, we define the posterior distribution as: $q_{\phi}(\mathbf{z}_{dia}|X_u, C_X, Y_u) \sim \mathcal{N}(\boldsymbol{\mu}'_{dia}, \boldsymbol{\sigma}'^2_{dia}\mathbf{I})$, where $\boldsymbol{\mu}'_{dia}$ and $\boldsymbol{\sigma}'_{dia}$ are calculated as:

$$\boldsymbol{\mu}'_{dia} = \mathrm{MLP}_{\phi}^{dia}(\mathbf{h}_X; \mathbf{h}_X^{ctx}; \mathbf{h}_Y),$$
$$\boldsymbol{\sigma}'_{dia} = \mathrm{Softplus}(\mathrm{MLP}_{\phi}^{dia}(\mathbf{h}_X; \mathbf{h}_X^{ctx}; \mathbf{h}_Y)).$$

Translation Consistency. To keep the lexical choice of translation consistent with those of previous utterances, we encode the paired source-target utterances and then sample a latent variable $\mathbf{z}_{tra} \in \mathbb{R}^{d_z}$. We define its prior distribution as: $p_{\theta}(\mathbf{z}_{tra}|X_u, C_X, C_Y) \sim \mathcal{N}(\boldsymbol{\mu}_{tra}, \boldsymbol{\sigma}^2_{tra}\mathbf{I})$ and $\{\boldsymbol{\mu}_{tra}, \boldsymbol{\sigma}_{tra}\}$ are calculated as:

$$\boldsymbol{\mu}_{tra} = \mathrm{MLP}_{\theta}^{tra}(\mathbf{h}_X; \mathbf{h}_X^{ctx}; \mathbf{h}_Y^{ctx}),$$
$$\boldsymbol{\sigma}_{tra} = \mathrm{Softplus}(\mathrm{MLP}_{\theta}^{tra}(\mathbf{h}_X; \mathbf{h}_X^{ctx}; \mathbf{h}_Y^{ctx})).$$

At training, the posterior distribution conditions on all paired bilingual dialogue utterances that contain implicit and aligned information, and serves as learning of the prior distribution. Specifically, we define the posterior distribution as: $q_{\phi}(\mathbf{z}_{tra}|X_u, C_X, C_Y, Y_u) \sim \mathcal{N}(\boldsymbol{\mu}'_{tra}, \boldsymbol{\sigma}'^2_{tra}\mathbf{I})$, where $\boldsymbol{\mu}'_{tra}$ and $\boldsymbol{\sigma}'_{tra}$ are calculated as:

$$\boldsymbol{\mu}'_{tra} = \mathrm{MLP}_{\phi}^{tra}(\mathbf{h}_X; \mathbf{h}_X^{ctx}; \mathbf{h}_Y^{ctx}; \mathbf{h}_Y),$$
$$\boldsymbol{\sigma}'_{tra} = \mathrm{Softplus}(\mathrm{MLP}_{\phi}^{tra}(\mathbf{h}_X; \mathbf{h}_X^{ctx}; \mathbf{h}_Y^{ctx}; \mathbf{h}_Y)).$$

4.4 Decoder

The decoder adopts a similar structure to the encoder, and each of N_d decoder layers contains an additional cross-attention sub-layer ($\mathrm{CrossAtt}$):

$$\mathbf{s}_d^{\ell} = \mathrm{SelfAtt}(\mathbf{h}_d^{\ell-1}) + \mathbf{h}_d^{\ell-1},\ \mathbf{h}_d^{\ell-1} \in \mathbb{R}^{n\times d},$$
$$\mathbf{c}_d^{\ell} = \mathrm{CrossAtt}(\mathbf{s}_d^{\ell}, \mathbf{h}_e^{N_e}) + \mathbf{s}_d^{\ell},\ \mathbf{s}_d^{\ell} \in \mathbb{R}^{n\times d},$$
$$\mathbf{h}_d^{\ell} = \mathrm{FFN}(\mathbf{c}_d^{\ell}) + \mathbf{c}_d^{\ell},\ \{\mathbf{c}_d^{\ell}, \mathbf{h}_d^{\ell}\} \in \mathbb{R}^{n\times d},$$

where $\mathbf{h}_d^{\ell}$ denotes the state of the ℓ-th decoder layer.

As shown in Fig. 3, we obtain the latent variables $\{\mathbf{z}_{role}, \mathbf{z}_{dia}, \mathbf{z}_{tra}\}$ either from the posterior distribution predicted by recognition networks (training process as the solid grey lines) or from prior distribution predicted by prior networks (inference process as the dashed red lines). Finally, we incorporate $\{\mathbf{z}_{role}, \mathbf{z}_{dia}, \mathbf{z}_{tra}\}$ into the state of the top layer of the decoder with a projection layer:

$$\mathbf{o}_t = \mathrm{Tanh}(\mathbf{W}_p[\mathbf{h}_{d,t}^{N_d}; \mathbf{z}_{role}; \mathbf{z}_{dia}; \mathbf{z}_{tra}] + \mathbf{b}_p),\ \mathbf{o}_t \in \mathbb{R}^d,$$

where $\mathbf{W}_p \in \mathbb{R}^{d\times(d+3d_z)}$ and $\mathbf{b}_p \in \mathbb{R}^d$ are training parameters, $\mathbf{h}_{d,t}^{N_d}$ is the hidden state at time-step t of the N_d-th decoder layer. Then, $\mathbf{o}_t$ is fed to a linear transformation and softmax layer to predict the probability distribution of the next target token:

$$\mathbf{p}_t = \mathrm{Softmax}(\mathbf{W}_o\mathbf{o}_t + \mathbf{b}_o),\ \mathbf{p}_t \in \mathbb{R}^{|V|},$$

where $\mathbf{W}_o \in \mathbb{R}^{|V|\times d}$ and $\mathbf{b}_o \in \mathbb{R}^{|V|}$ are training parameters.

4.5 Training Objectives

We apply a two-stage training strategy (Zhang et al., 2018; Ma et al., 2020). Firstly, we train our model on large-scale sentence-level NMT data to minimize the cross-entropy objective:

$$\mathcal{L}(\theta; X, Y) = -\sum_{t=1}^{N} \log p_{\theta}(y_t|X, y_{1:t-1}).$$

Secondly, we fine-tune it on the chat translation data to maximize the following objective:

$$\begin{aligned}
&\mathcal{J}(\theta, \phi; X_u, C_X^{role}, C_X, C_Y, Y_u) = \\
&-\mathrm{KL}(q_{\phi}(\mathbf{z}_{role}|X_u, C_X^{role}, Y_u)\|p_{\theta}(\mathbf{z}_{role}|X_u, C_X^{role})) \\
&-\mathrm{KL}(q_{\phi}(\mathbf{z}_{dia}|X_u, C_X, Y_u)\|p_{\theta}(\mathbf{z}_{dia}|X_u, C_X)) \\
&-\mathrm{KL}(q_{\phi}(\mathbf{z}_{tra}|X_u, C_X, C_Y, Y_u)\|p_{\theta}(\mathbf{z}_{tra}|X_u, C_X, C_Y)) \\
&+\mathbb{E}_{q_{\phi}}[\log p_{\theta}(Y_u|X_u, \mathbf{z}_{role}, \mathbf{z}_{dia}, \mathbf{z}_{tra})].
\end{aligned}$$

We use the reparameterization trick (Kingma and Welling, 2013) to estimate the gradients of the prior and recognition networks (Zhao et al., 2017).

5 Experiments

5.1 Datasets and Metrics

Datasets. We apply a two-stage training strategy, *i.e.*, firstly training on a large-scale sentence-level NMT corpus (WMT20[6]) and then fine-tuning on chat translation corpus (BConTrasT (Farajian et al., 2020)[7] and BMELD). The details (WMT20 data and results of the first stage) are shown in Appendix A.

BConTrasT. The dataset[8] is first provided by WMT 2020 Chat Translation Task (Farajian et al., 2020), which is translated from English into German and is based on the monolingual Taskmaster-1 corpus (Byrne et al., 2019). The conversations (originally in English) were first automatically translated into German and then manually post-edited by Unbabel editors,[9] who are native German speakers. Having the conversations in both languages allows us to simulate bilingual conversations in which one speaker, the customer, speaks in German and the other speaker, the agent, answers in English.

BMELD. Similarly, based on the dialogue dataset in the MELD (originally in English) (Poria et al., 2019),[10] we firstly crawled the corresponding Chinese translations from this[11] and then manually post-edited them according to the dialogue history by native Chinese speakers, who are postgraduate students majoring in English. Finally, following (Farajian et al., 2020), we assume 50% speakers as Chinese speakers to keep data balance for Ch⇒En translations and build the bilingual MELD (BMELD). For the Chinese, we segment the sentence using Stanford CoreNLP toolkit[12].

Metrics. For fair comparison, we use the SacreBLEU[13] (Post, 2018) and v0.7.25 for TER (Snover et al., 2006) (the lower the better) with the statistical significance test (Koehn, 2004). For En⇔De, we report case-sensitive score following the WMT20 chat task (Farajian et al., 2020). For Ch⇒En, we report case-insensitive score. For En⇒Ch, we report the character-level BLEU score.

Dataset	# Dialogues			# Utterances		
	Train	Valid	Test	Train	Valid	Test
En⇒De	550	78	78	7,629	1,040	1,133
De⇒En	550	78	78	6,216	862	967
En⇒Ch	1,036	108	274	5,560	567	1,466
Ch⇒En	1,036	108	274	4,427	517	1,135

Table 1: Statistics of chat translation data.

5.2 Implementation Details

For all experiments, we follow the *Transformer-Base* and *Transformer-Big* settings illustrated in (Vaswani et al., 2017). In *Transformer-Base*, we use 512 as hidden size (*i.e.*, d), 2048 as filter size and 8 heads in multi-head attention. In *Transformer-Big*, we use 1024 as hidden size, 4096 as filter size, and 16 heads in multi-head attention. All our Transformer models contain $N_e = 6$ encoder layers and $N_d = 6$ decoder layers and all models are trained using THUMT (Tan et al., 2020) framework. We conduct experiments on the validation set of En⇒De to select the hyperparameters of context length and latent dimension, which are then shared for all tasks. For the results and more details (other hyperparameters setting and average running time), please refer to Appendix B, C, and D.

5.3 Comparison Models

Baseline NMT Models. Transformer (Vaswani et al., 2017): the de-facto NMT model that does not fine-tune on chat translation data. Transformer+FT: fine-tuning on the chat translation data after being pre-trained on sentence-level NMT corpus.

Context-Aware NMT Models. Doc-Transformer+FT (Ma et al., 2020): a state-of-the-art document-level NMT model based on Transformer sharing the first encoder layer to incorporate the bilingual dialogue history. Dia-Transformer+FT (Maruf et al., 2018): using an additional RNN-based (Hochreiter and Schmidhuber, 1997) encoder to incorporate the mixed-language dialogue history, where we re-implement it based on Transformer and use another Transformer layer to introduce context. V-Transformer+FT (Zhang et al., 2016; McCarthy

[6] http://www.statmt.org/wmt20/translation-task.html
[7] http://www.statmt.org/wmt20/chat-task.html
[8] https://github.com/Unbabel/BConTrasT
[9] www.unbabel.com
[10] The MELD is a multimodal emotionLines dialogue dataset, each utterance of which corresponds to a video, voice, and text, and is annotated with detailed emotion and sentiment.
[11] https://www.zimutiantang.com/
[12] https://stanfordnlp.github.io/CoreNLP/index.html
[13] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.13

	Models	En⇒De		De⇒En		En⇒Ch		Ch⇒En	
		BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓
Baseline NMT models (Base)	Transformer	40.02	42.5	48.38	33.4	21.40	72.4	18.52	59.1
	Transformer+FT	58.43	26.7	<u>59.57</u>	26.2	25.22	62.8	21.59	56.7
Context-Aware NMT models (Base)	Doc-Transformer+FT	58.15	27.1	59.46	<u>25.7</u>	24.76	63.4	20.61	59.8
	Dia-Transformer+FT	58.33	26.8	59.09	26.2	24.96	63.7	20.49	60.1
	V-Transformer+FT	<u>58.74</u>	<u>26.3</u>	58.67	27.0	<u>26.82</u>	<u>60.6</u>	<u>21.86</u>	<u>56.3</u>
Ours (Base)	CPCC	**60.13**††	**25.4**††	**61.05**††	**24.9**††	**27.55**†	**60.1**†	**22.50**†	**55.7**†
Baseline NMT models (Big)	Transformer	40.53	42.2	49.90	33.3	22.81	69.6	19.58	57.7
	Transformer+FT	<u>59.01</u>	<u>26.0</u>	59.98	25.9	26.95	60.7	22.15	56.1
Context-Aware NMT models (Big)	Doc-Transformer+FT	58.61	26.5	59.98	<u>25.4</u>	26.45	62.6	21.38	57.7
	Dia-Transformer+FT	58.68	26.8	59.63	26.0	26.72	62.4	21.09	58.1
	V-Transformer+FT	58.70	26.2	<u>60.01</u>	25.7	<u>27.52</u>	<u>60.3</u>	<u>22.24</u>	<u>55.9</u>
Ours (Big)	CPCC	**60.23**††	**25.6**†	**61.45**††	**24.8**†	**28.98**††	**59.0**††	**22.98**†	**54.6**††

Table 2: Results on BConTrasT (En⇔De) and BMELD (En⇔Ch) in terms of BLEU (%) and TER (%). The best and the second results are bold and underlined, respectively. "†" and "††" indicate that statistically significant better than the best result of all contrast NMT models with t-test $p < 0.05$ and $p < 0.01$, respectively.

#	Models	En⇒De		De⇒En	
		BLEU↑	TER↓	BLEU↑	TER↓
0	CPCC (*Base*)	**60.96**	**24.6**	**62.09**	**24.5**
1	*w/o* $\mathbf{z}_{role}$	60.56 (-0.40)	25.1	61.42 (-0.67)	24.8
2	*w/o* $\mathbf{z}_{dia}$	60.50 (-0.46)	25.2	61.65 (-0.44)	25.1
3	*w/o* $\mathbf{z}_{tra}$	60.39 (-0.57)	25.1	61.38 (-0.71)	26.0
4	*w/o* $\mathbf{z}_{role}$ & $\mathbf{z}_{dia}$	59.64 (-1.32)	25.8	60.65 (-1.44)	25.8
5	*w/o* $\mathbf{z}_{role}$ & $\mathbf{z}_{tra}$	59.61 (-1.35)	25.9	60.62 (-1.47)	25.7
6	*w/o* $\mathbf{z}_{dia}$ & $\mathbf{z}_{tra}$	60.24 (-0.72)	25.1	61.18 (-0.91)	24.9
7	*w/o* all	58.95 (-2.01)	26.1	59.82 (-2.27)	26.1

Table 3: Ablation study on the validation set. "*w/o* all" indicates removing all latent variables but remaining encoding all bilingual dialogue history.

et al., 2020): the variational NMT model based on Transformer also sharing the first encoder layer to exploit the bilingual context for fair comparison.

5.4 Main Results

Overall, we separate the models into two parts in Tab. 2: the *Base* setting and the *Big* setting. In each part, we show the results of our re-implemented Transformer baselines, the context-aware NMT systems, and our approach on En⇔De and En⇔Ch.

Results on En⇔De. Under the *Base* setting, CPCC substantially outperforms the baselines (*e.g.*, "Transformer+FT") by a large margin with 1.70↑ and 1.48↑ BLEU scores on En⇒De and De⇒En, respectively. On the TER, our CPCC achieves a significant improvement of 1.3 points in both language pairs. Under the *Big* setting, our CPCC also consistently boosts the performance in both directions (*i.e.*, 1.22↑ and 1.47↑ BLEU scores, 0.4↓ and 1.1↓ TER scores), showing its effectiveness.

Compared against the strong context-aware NMT systems (underlined results), our CPCC significantly surpasses them (about 1.39∼1.59↑ BLEU scores and 0.6∼0.9↓ TER scores) in both language directions under both *Base* and *Big* settings, demonstrating the superiority of our model.

Results on En⇔Ch. We also conduct experiments on our self-collected data to validate the generalizability across languages in Tab. 2.

Our CPCC presents remarkable BLEU improvements over the "Transformer+FT" by a large margin in two directions by 2.33↑ and 0.91↑ BLEU gains under the *Base* setting, respectively, and by 2.03↑ and 0.83↑ BLEU gains in both directions under the *Big* setting. These results suggest that CPCC consistently performs well across languages.

Compared with strong context-aware NMT systems (*e.g.*, "V-Transformer+FT"), our approach notably surpasses them in both language directions under both *Base* and *Big* settings, which shows the generalizability and superiority of our model.

6 Analysis

6.1 Ablation Study

We conduct ablation studies to investigate how well each tailored latent variable of our model works. When removing latent variables listed in Tab. 3, we have the following findings.

(1) All latent variables make substantial contributions to performance, proving the importance of modeling role preference, dialogue coherence, and translation consistency, which is consistent with our intuition that the properties should be beneficial to better translations (rows 1∼3 vs. row 0).

(2) Results of rows 4∼7 show the combination effect of three latent variables, suggesting that the combination among three latent variables has a cumulative effect (rows 4∼7 vs. rows 0∼3).

(3) Row 7 vs. row 0 shows that explicitly modeling the bilingual conversational characteristics significantly outperforms implicit modeling (*i.e.*, just incorporating the dialogue history into encoders), which lacks the relevant information guidance.

6.2 Dialogue Coherence

Following (Lapata and Barzilay, 2005; Xiong et al., 2019), we measure dialogue coherence as sentence similarity. Specifically, the representation of each sentence is the mean of the distributed vectors of its words, and the dialogue coherence between two sentences s_1 and s_2 is determined by the cosine similarity:

$$sim(s_1, s_2) = cos(f(s_1), f(s_2)),$$
$$f(s_i) = \frac{1}{|s_i|} \sum_{\mathbf{w} \in s_i} (\mathbf{w}),$$

where $\mathbf{w}$ is the vector for word w.

We use Word2Vec[14] (Mikolov et al., 2013) to learn the distributed vectors of words by training on the monolingual dialogue dataset: Taskmaster-1 (Byrne et al., 2019). And we set the dimensionality of word embeddings to 100.

Tab. 4 shows the cosine similarity on the test set of De⇒En. It reveals that our model encouraged by tailor-made latent variables produces better coherence in chat translation than contrast systems.

6.3 Human Evaluation

Inspired by (Bao et al., 2020; Farajian et al., 2020), we use four criteria for human evaluation: (1) **Preference** measures whether the translation preserves the role preference information; (2) **Coherence** denotes whether the translation is semantically coherent with the dialogue history; (3) **Consistency** measures whether the lexical choice of translation is consistent with the preceding utterances; (4) **Fluency** measures whether the translation is logically reasonable and grammatically correct.

[14]https://code.google.com/archive/p/word2vec/

Models	1-th Pr.	2-th Pr.	3-th Pr.
Transformer	0.6502	0.6037	0.5659
Transformer+FT	0.6587	0.6104	0.5714
Doc-Transformer+FT	0.6569	0.6093	0.5713
Dia-Transformer+FT	0.6553	0.6084	0.5709
V-Transformer+FT	0.6602	0.6122	0.5751
CPCC (Ours)	0.6660††	**0.6190**††	**0.5814**††
Human Reference	**0.6663**	**0.6190**	0.5795

Table 4: Results of dialogue coherence in terms of sentence similarity (De⇒En, *Base*). The "**#-th Pr.**" denotes the #-th preceding utterance to the current one. "††" indicates that statistically significant better than the best result of all contrast NMT models ($p < 0.01$).

Models	Pref.	Coh.	Con.	Flu.
Transformer	0.485	0.540	0.510	0.590
Transformer+FT	0.530	0.590	0.565	0.635
Doc-Transformer+FT	0.525	0.595	0.560	0.630
Dia-Transformer+FT	0.525	0.580	0.555	0.625
V-Transformer+FT	0.535	0.595	0.560	0.635
CPCC (Ours)	**0.570**	**0.620**	**0.585**	**0.650**

Table 5: Results of Human evaluation (Ch⇒En, *Base*). "**Pref.**": Preference. "**Coh.**": Coherence. "**Con.**": Consistency. "**Flu.**": Fluency.

We firstly randomly sample 200 examples from the test set of Ch⇒En. Then, we assign each bilingual dialogue history and corresponding 6 generated translations to three human annotators without order, and ask them to evaluate whether each translation meets the criteria defined above. All annotators are postgraduate students and not involved in other parts of our experiments.

Tab. 5 shows that our CPCC effectively alleviates the problem of role-irrelevant, incoherent and inconsistent translations compared with other models (significance test (Koehn, 2004), $p < 0.05$), indicating the superiority of our model. The inter-annotator agreement is 0.527, 0.491, 0.556 and 0.485 calculated by the Fleiss' kappa (Fleiss and Cohen, 1973), for preference, coherence, consistency and fluency, respectively, indicating "Moderate Agreement" for all four criteria. We also present some case studies in Appendix H.

7 Related Work

Chat NMT. It only involves several researches due to the lack of human-annotated publicly available data (Farajian et al., 2020). Therefore, some

existing work (Wang et al., 2016; Maruf et al., 2018; Zhang and Zhou, 2019; Rikters et al., 2020) mainly pays attention to designing methods to automatically construct the subtitles corpus, which may contain noisy bilingual utterances. Recently, Farajian et al. (2020) organize the WMT20 chat translation task and first provide a human post-edited corpus, where some teams investigate the effect of dialogue history and finally ensemble their models for higher ranks (Berard et al., 2020; Mohammed et al., 2020; Wang et al., 2020; Bao et al., 2020; Moghe et al., 2020). As a synchronizing study, Wang et al. (2021) use multitask learning to auto-correct the translation error, such as pronoun dropping, punctuation dropping, and typos. Unlike them, we focus on explicitly modeling role preference, dialogue coherence, and translation consistency with tailored latent variables to promote the translation quality.

Context-Aware NMT. Chat NMT can be viewed as a special case of context-aware NMT, which has attracted many researchers (Gong et al., 2011; Jean et al., 2017; Wang et al., 2017b; Bawden et al., 2018; Miculicich et al., 2018; Kuang et al., 2018; Tu et al., 2018; Yang et al., 2019; Kang et al., 2020; Li et al., 2020; Ma et al., 2020) to extend the encoder or decoder for exploring the context impact on translation quality. Although these models can be directly applied to chat translation, they cannot explicitly capture the bilingual conversational characteristics and thus lead to unsatisfactory translations (Moghe et al., 2020). Different from these studies, we focus on explicitly modeling these bilingual conversational characteristics via CVAE for better translations.

Conditional Variational Auto-Encoder. CVAE has verified its superiority in many fields (Sohn et al., 2015). In NMT, Zhang et al. (2016) and Su et al. (2018) extend CVAE to capture the global/local information of source sentence for better results. McCarthy et al. (2020) focus on addressing the posterior collapse with mutual information. Besides, some studies use CVAE to model the correlations between image and text for multimodal NMT (Toyama et al., 2016; Calixto et al., 2019). Although the CVAE has been widely used in NLP tasks, its adaption and utilization to chat translation for modeling inherent bilingual conversational characteristics are non-trivial, and to the best of our knowledge, has never been investigated before.

8 Conclusion and Future Work

We propose to model bilingual conversational characteristics through tailored latent variables for neural chat translation. Experiments on En⇔De and En⇔Ch directions show that our model notably improves translation quality on both BLEU and TER metrics, showing its superiority and generalizability. Human evaluation further verifies that our model yields role-specific, coherent, and consistent translations by incorporating tailored latent variables into NMT. Moreover, we contribute a new bilingual dialogue data (BMELD, En⇔Ch) with manual translations to the research community. In the future, we would like to explore the effect of multimodality and emotion on chat translation, which has been well studied in dialogue field (Liang et al., 2020).

Acknowledgments

The research work descried in this paper has been supported by the National Key R&D Program of China (2020AAA0108001) and the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

JinYeong Bak and Alice Oh. 2019. Variational hierarchical user-based conversation model. In *Proceedings of EMNLP-IJCNLP*, pages 1941–1950.

Calvin Bao, Yow-Ting Shiue, Chujun Song, Jie Li, and Marine Carpuat. 2020. The university of maryland's submissions to the wmt20 chat translation task: Searching for more data to adapt discourse-aware neural machine translation. In *Proceedings of WMT*, pages 454–459.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of NAACL*, pages 1304–1313.

Alexandre Berard, Ioan Calapodescu, Vassilina Nikoulina, and Jerin Philip. 2020. Naver labs europe's participation in the robustness, chat, and biomedical tasks at wmt 2020. In *Proceedings of WMT*, pages 460–470.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel

Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of EMNLP-IJCNLP*, pages 4516–4525.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of ACL*, pages 6392–6405.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of WMT*, pages 65–75.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, pages 613–619.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of EMNLP*, pages 909–919.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–1780.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of EMNLP*, pages 2242–2254.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of COLING*, pages 596–606.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of IJCAI*, pages 1085–1090.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of EMNLP*, pages 4791–4796.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of ACL*, pages 3512–3518.

Yunlong Liang, Fandong Meng, Ying Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2020. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of ACL*, pages 3505–3511.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of ACL*, pages 1275–1284.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of WMT*, pages 101–112.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of NAACL*, pages 3092–3102.

Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong. 2020. Addressing posterior collapse with mutual information for improved variational neural machine translation. In *Proceedings of ACL*, pages 8512–8525.

Fandong Meng and Jinchao Zhang. 2019. DTMT: A novel deep transition architecture for neural machine translation. In *Proceedings of AAAI*, pages 224–231.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of EMNLP*, pages 2947–2954.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of EMNLP*, pages 1102–1108.

Nikita Moghe, Christian Hardmeier, and Rachel Bawden. 2020. The university of edinburgh-uppsala university's submission to the wmt 2020 chat translation task. In *Proceedings of WMT*, pages 471–476.

Roweida Mohammed, Mahmoud Al-Ayyoub, and Malak Abdullah. 2020. Just system for wmt20 chat translation task. In *Proceedings of WMT*, pages 477–480.

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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*, pages 527–536.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*, pages 186–191.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2020. Document-aligned Japanese-English conversation parallel corpus. In *Proceedings of MT*, pages 639–645, Online.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Proceedings of NIPS*, pages 3483–3491.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. In *Proceedings of AAAI*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.

Zhixing Tan, Jiacheng Zhang, Xuancheng Huang, Gang Chen, Shuo Wang, Maosong Sun, Huanbo Luan, and Yang Liu. 2020. THUMT: An open-source toolkit for neural machine translation. In *Proceedings of AMTA*, pages 116–122.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the DiscoMT*, pages 82–92.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of WMT*, pages 113–123.

Joji Toyama, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Neural machine translation with latent semantic of image and text. *arXiv preprint arXiv:1611.08459*.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *TACL*, pages 407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of EMNLP-IJCNLP*, pages 877–886.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of ACL*, pages 1198–1212.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, pages 1264–1274.

Longyue Wang, Jinhua Du, Liangyou Li, Zhaopeng Tu, Andy Way, and Qun Liu. 2017a. Semantics-enhanced task-oriented dialogue translation: A case study on hotel booking. In *Proceedings of IJCNLP*, pages 33–36.

Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020. Tencent ai lab machine translation systems for wmt20 chat translation task. In *Proceedings of WMT*, pages 481–489.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One model to learn both: Zero pronoun prediction and translation. In *Proceedings of EMNLP-IJCNLP*, pages 921–930.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017b. Exploiting cross-sentence context for neural machine translation. In *Proceedings of EMNLP*, pages 2826–2831.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016. Automatic construction of discourse corpora for dialogue translation. In *Proceedings of LREC*, pages 2748–2754.

Tao Wang, Chengqi Zhao, Mingxuan Wang, Lei Li, and Deyi Xiong. 2021. Autocorrect in the process of translation – multi-task learning improves dialogue machine translation.

Tianming Wang and Xiaojun Wan. 2019. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *Proceedings of IJCAI*, pages 5233–5239.

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *Proceedings of ACL*, pages 53–65.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. *Proceedings of AAAI*, pages 7338–7345.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2020. Multi-unit transformers for neural machine translation. In *Proceedings of EMNLP*, pages 1047–1059, Online.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings of ACL*, pages 2002–2012.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of EMNLP*, pages 521–530.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of EMNLP*, pages 533–542.

L. Zhang and Q. Zhou. 2019. Automatically annotate tv series subtitles for dialogue corpus construction. In *APSIPA ASC*, pages 1029–1035.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of ACL*, pages 4334–4343, Florence, Italy.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of ACL*, pages 654–664.

Appendix

A Datasets

WMT20. For the En⇔De, we combine six corpora including Euporal, ParaCrawl, CommonCrawl, TildeRapid, NewsCommentary, and WikiMatrix, and we combine News Commentary v15, Wiki Titles v2, UN Parallel Corpus V1.0, CCMT Corpus, and WikiMatrix for the En⇔Ch. We firstly filter noisy sentence pairs according to their characteristics in terms of duplication and length (whose length exceeds 80). To pre-process the raw data, we employ a series of open-source/in-house scripts, including full-/half-width conversion, unicode conversation, punctuation normalization, and tokenization (Wang et al., 2020). After filtering steps, we generate subwords via joint BPE (Sennrich et al., 2016) with 32K merge operations. Finally, we obtain 45,541,367 sentence pairs for En⇔De and 22,244,006 sentence pairs for En⇔Ch, respectively.

We test the model performance of the first stage on *newstest2019*. The results are shown in Tab. 6.

	Methods	**En⇒De**	**De⇒En**	**En⇒Ch**	**Ch⇒En**
Base	Transformer	39.88	40.72	32.55	24.42
	V-Transformer	40.01	41.36	32.90	25.77
Big	Transformer	41.35	41.56	33.85	24.86
	V-Transformer	41.40	41.67	33.90	26.46

Table 6: The BLEU scores on the *newstest2019* of the first stage.

B Implementation Details

For all experiments, we follow two model settings illustrated in (Vaswani et al., 2017), namely *Transformer-Base* and *Transformer-Big*. The training step is set to 200,000 and 2,000 for the first stage and the fine-tuning stage, respectively. The batch size for each GPU is set to 4096 tokens. The beam size is set to 4, and the length penalty is 0.6 among all experiments. All experiments in the first stage are conducted utilizing 8 NVIDIA Tesla V100 GPUs, while we use 2 GPUs for the second stage, *i.e.*, fine-tuning. That gives us about 8*4096 and 2*4096 tokens per update for all experiments in the first-stage and second-stage, respectively. All models are optimized using Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.998$, and learning rate is set to 1.0 for all experiments. Label smoothing is set to 0.1. We use dropout of 0.1/0.3 for *Base* and *Big* setting, respectively. To alleviate the degeneration problem of the variational framework, we apply KL annealing. The KL multiplier λ gradually increases from 0 to 1 over 10, 000 steps. $|R|$ is set to 2 for En⇔De and 7 for En⇔Ch, respectively. $|T|$ is set to 10. The criterion for selecting hyperparameters is the BLEU score on validation sets for both tasks. The average running time is shown in Tab. 7.

	Stages	En⇒De	De⇒En	En⇒Ch	Ch⇒En
Base	The First Stage	5D	7D	4D	3.5D
	Fine-Tuning Stage	4H	5H	3H	2H
Big	The First Stage	10D	12D	7D	6D
	Fine-Tuning Stage	4.5H	5.5H	4H	2.5H

Table 7: The average running time for the first stage and fine-tuning stage. D: Days, H: Hours.

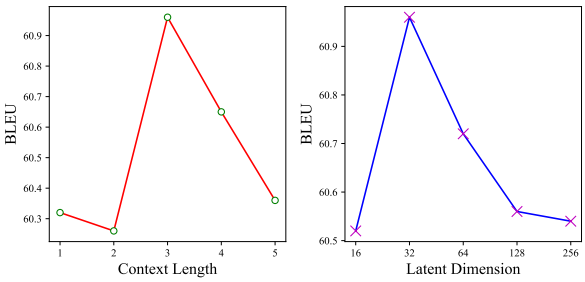

Figure 4: Effect of context length and latent dimension on translation quality. The BLEU scores (%) are calculated on the validation set of the En⇒De.

In the case of blind testing or online use (assumed dealing with En⇒De), since translations of target utterances (*i.e.*, English) will not be given, an inverse De⇒En model is simultaneously trained and used to back-translate target utterances (Bao et al., 2020), similar to all tasks.

C Effect of Context Length

We firstly investigate the effect of context length (*i.e.*, the number of preceding utterances) on our approach under the Transformer *Base* setting. As shown in the left of Fig. 4, using three preceding source sentences as dialogue history achieves the best translation performance on the validation set (En⇒De). Using more preceding sentences does not bring any improvement and increases the computational cost. This confirms the finding of Tu et al. (2018) and Zhang et al. (2018) that long-distance context only has limited influence. Therefore, we set the number of preceding sentences to 3 in all experiments.

D Effect of Latent Dimension

The right of Fig. 4 shows the effect of the latent dimension on translation quality under the Transformer *Base* setting. Obviously, using latent dimension 32 suffices to achieve superior performance. Increasing the dimension does not lead to any improvements. Therefore, we set the latent dimension to 32 in all experiments.

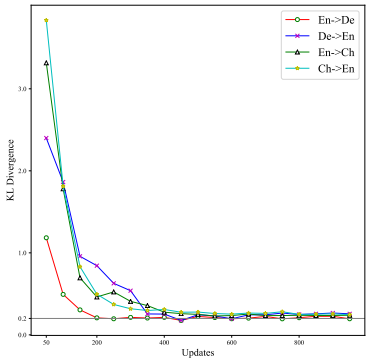

Figure 5: Total KL divergence (per word) of all latent variables (first 1,000 updates on corresponding validation set).

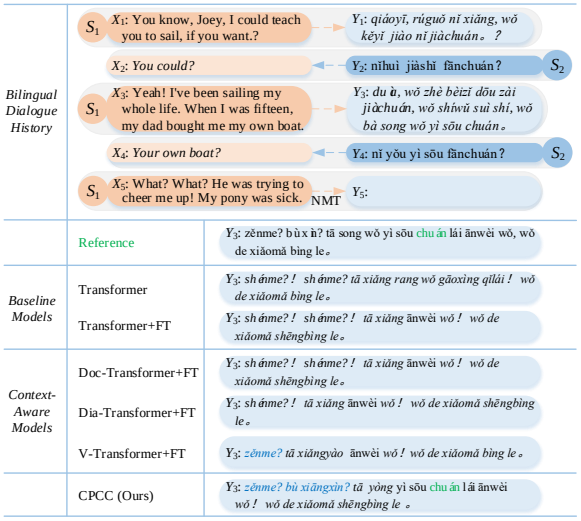

Figure 6: Bilingual conversational example one.

E KL Divergence

Generally, KL divergence measures the amount of information encoded in a latent variable. In the extreme case where the KL divergence of latent variable $\mathbf{z}$ equals to zero, the model completely ignores $\mathbf{z}$, *i.e.*, it degenerates. Fig. 5 shows that the total KL divergence of our model maintains around 0.2∼0.5 indicating that the degeneration problem does not exist in our model and latent variables can play their corresponding roles.

F Case Study

In this section, we show some cases in Fig. 6 and Fig. 7 to investigate the effect of different models.

Role Preference and Dialogue Coherence. As shown in Fig. 6, we observe that the baseline models and the context-aware models except "V-Transformer+FT" cannot preserve the role preference information, *e.g.*, *joy* emotion, even these

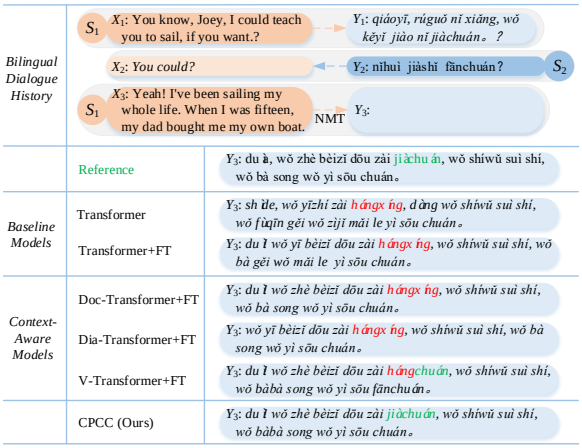

Figure 7: Bilingual conversational example two.

"*-Transformer+FT" models incorporate the bilingual conversational history into the encoder. The "V-Transformer+FT" model produces very slightly emotional elements (*e.g.*, "*zěnme?*") due to the latent variable over the source sentence capturing relevant preference information. Meanwhile, we find that all comparison models cannot generate a coherent translation. The reason may be that they fail to capture the conversation-level coherence clue, *i.e.*, "*boat*". By contrast, we explicitly model the two characteristics through tailored latent variables and thus obtain satisfactory results.

Translation Consistency. As shown in Fig. 7, we observe that all comparison models cannot maintain the translation consistency due to the lack of explicitly modeling this characteristic. Our model has the ability to overcome the issue and can keep the correct lexical choice to translate the current utterance that might have appeared in preceding turns, *i.e.*, "*jiàchuàn*".

To sum up, both cases show that our model yields role-specific, coherent, and consistent translations by incorporating tailored latent variables into translators, demonstrating its effectiveness and superiority.