Towards Making the Most of Dialogue Characteristics for Neural Chat Translation

Yunlong Liang[1]*, **Chulun Zhou**[2]*, **Fandong Meng**[3], **Jinan Xu**[1]†,
Yufeng Chen[1], **Jinsong Su**[2] and **Jie Zhou**[3]
[1]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University
[2]School of Informatics, Xiamen University
[3]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{yunlongliang,chenyf,jaxu}@bjtu.edu.cn clzhou@stu.xmu.edu.cn
jssu@xmu.edu.cn {fandongmeng,withtomzhou}@tencent.com

Abstract

Neural Chat Translation (NCT) aims to translate conversational text between speakers of different languages. Despite the promising performance of sentence-level and context-aware neural machine translation models, there still remain limitations in current NCT models because the inherent dialogue characteristics of chat, such as dialogue coherence and speaker personality, are neglected. In this paper, we propose to promote the chat translation by introducing the modeling of dialogue characteristics into the NCT model. To this end, we design four auxiliary tasks including *monolingual response generation*, *cross-lingual response generation*, *next utterance discrimination*, and *speaker identification*. Together with the main chat translation task, we optimize the NCT model through the training objectives of all these tasks. By this means, the NCT model can be enhanced by capturing the inherent dialogue characteristics, thus generating more coherent and speaker-relevant translations. Comprehensive experiments on four language directions (English⇔German and English⇔Chinese) verify the effectiveness and superiority of the proposed approach.

1 Introduction

A cross-lingual conversation involves participants that speak in different languages (*e.g.*, one speaking in English and another in Chinese as shown in Fig. 1), where a chat translator can be applied to help participants communicate in their individual native languages. The chat translator converts the language of bilingual conversational text in both directions, *e.g.* from English to Chinese and vice versa (Farajian et al., 2020). With more international communication worldwide, the chat translation task becomes more important and has a wider range of applications.

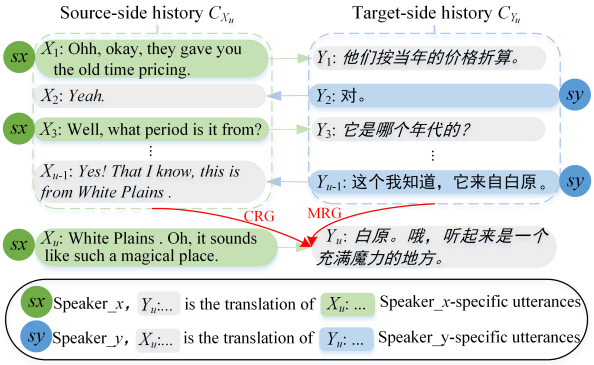

Figure 1: A dialogue example (En⇔Zh) when translating the utterance X_u. CRG: cross-lingual response generation. MRG: monolingual response generation.

In recent years, although sentence-level Neural Machine Translation (NMT) models (Sutskever et al., 2014; Vaswani et al., 2017; Hassan et al., 2018; Meng and Zhang, 2019; Yan et al., 2020; Zhang et al., 2019) have achieved remarkable progress and can be directly used as the chat translator, they often lead to incoherent and speaker-irrelevant translations (Mirkin et al., 2015; Wang et al., 2017a; Läubli et al., 2018; Toral et al., 2018) due to ignoring the chat history that contains useful contextual information. To exploit chat history, context-aware NMT models (Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Bawden et al., 2018; Miculicich et al., 2018; Tu et al., 2018; Voita et al., 2018, 2019a,b; Wang et al., 2019a; Maruf et al., 2019; Chen et al., 2020; Ma et al., 2020, etc) can also be directly adapted to chat translation. However, their performances are usually limited because of lacking the modeling of the inherent dialogue characteristics (*e.g.*, the dialogue coherence and speaker personality), which matter for chat translation task as pointed out by Farajian et al. (2020).

In this paper, we propose a **C**oherence-**S**peaker-

* Equal contribution. Work was done when Liang and Zhou were interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.

† Jinan Xu is the corresponding author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 67–79
November 7–11, 2021.

Aware NCT (CSA-NCT) training framework to improve the NCT model by making use of dialogue characteristics in conversations. Concretely, from the perspectives of dialogue coherence and speaker personality, we design four auxiliary tasks along with the main chat translation task. For dialogue coherence, there are three tasks (two generation tasks and one discrimination task), namely *monolingual response generation*, *cross-lingual response generation*, and *next utterance discrimination*. Specifically, as shown in Fig. 1, (1) the monolingual response generation task aims to generate the coherent corresponding utterance in target language given the dialogue history context of the same language. Similarly, (2) the cross-lingual response generation task is to leverage the dialogue history context in source language to generate the coherent corresponding utterance in target language. Besides the above two generation tasks, (3) the next utterance discrimination task focuses on distinguishing whether the translated text is coherent to be the next utterance of the given dialogue history context. Moreover, for speaker personality, (4) we design the speaker identification task that judges whether the translated text is consistent with the personality of its original speaker. Together with the main chat translation task, the NCT model is optimized through the joint objectives of all these auxiliary tasks. In this way, the model is enhanced to capture dialogue coherence and speaker personality in conversation, which thus can generate more coherent and speaker-relevant translations.

We validate our CSA-NCT framework on the datasets of different language pairs: BConTrasT (Farajian et al., 2020) (En⇔De[1]) and BMELD (Liang et al., 2021a) (En⇔Zh[2]). The experimental results show that our model achieves consistent improvements on four translation tasks in terms of both BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), demonstrating its effectiveness and generalizability. Human evaluation further suggests that our model can generate more coherent and speaker-relevant translations compared to the existing related methods.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to incorporate the dialogue coherence and speaker personality into neural chat translation.
- We propose a multi-task learning framework with four auxiliary tasks to help the NCT model generate more coherent and speaker-relevant translations.
- Extensive experiments on datasets of different language pairs demonstrate that our model with multi-task learning achieves the state-of-the-art performances on the chat translation task and significantly outperforms the existing sentence-level/context-aware NMT models.[3]

2 Background

Sentence-Level NMT. Given an input sequence $X=\{x_i\}_{i=1}^{|X|}$, the goal of the sentence-level NMT model is to generate its translation $Y=\{y_i\}_{i=1}^{|Y|}$. The model is optimized through the following objective:

$$\mathcal{L}_{\text{S-NMT}} = -\sum_{t=1}^{|Y|} \log(p(y_t|X, y_{<t})). \quad (1)$$

Context-Aware NMT. As in (Ma et al., 2020), given a paragraph of input sentences $D^X = \{X_j\}_{j=1}^{J}$ in source language and its corresponding translations $D^Y = \{Y_j\}_{j=1}^{J}$ in target language with J paired sentences, the training objective of a context-aware NMT model can be formalized as

$$\mathcal{L}_{\text{C-NMT}} = -\sum_{j=1}^{J} \log(p(Y_j|X_j, X_{<j}, Y_{<j})), \quad (2)$$

where $X_{<j}$ and $Y_{<j}$ are the preceding contexts of the j-th input source sentence and the j-th target translation, respectively.

3 CSA-NCT Training Framework

In this section, we introduce the proposed CSA-NCT training framework, which aims to improve the NCT model with four elaborately designed auxiliary tasks. In the following subsections, we first describe the problem formalization (§ 3.1) and the NCT model (§ 3.2). Then, we introduce each auxiliary task in detail (§ 3.3). Finally, we elaborate the process of training and inference (§ 3.4).

3.1 Problem Formalization

In the scenario of this paper, the chat involves two speakers (sx and sy) speaking in two languages. As shown in Fig. 1, we assume the

[1]En⇔De: English⇔German.
[2]En⇔Zh: English⇔Chinese.
[3]The code is publicly available at: https://github.com/XL2248/CSA-NCT

two speakers have alternately given utterances in their individual languages for u turns, resulting in $X_1, X_2, X_3, X_4, X_5, ..., X_{u-1}, X_u$ and $Y_1, Y_2, Y_3, Y_4, Y_5, ..., Y_{u-1}, Y_u$ on the source and target sides, respectively. Among these utterances, $X_1, X_3, X_5, ..., X_u$ are originally spoken by the speaker sx and $Y_1, Y_3, Y_5, ..., Y_u$ are the corresponding translations in target language. Analogously, $Y_2, Y_4, Y_6, ..., Y_{u-1}$ are originally spoken by the speaker sy and $X_2, X_4, X_6, ..., X_{u-1}$ are the translated utterances in source language.

According to languages, we define the dialogue history context of X_u on the source side as $\mathcal{C}_{X_u}$=$\{X_1, X_2, X_3, X_4, X_5, ..., X_{u-1}\}$ and that of Y_u on the target side as C_{Y_u}=$\{Y_1, Y_2, Y_3, Y_4, Y_5, ..., Y_{u-1}\}$. According to original speakers, on the target side, we define the speaker sx-specific dialogue history context of Y_u as the partial set of its preceding utterances $\mathcal{C}^{sx}_{Y_u}$=$\{Y_1, Y_3, Y_5, ..., Y_{u-2}\}$ and the speaker sy-specific dialogue history context of Y_u as $\mathcal{C}^{sy}_{Y_u}$=$\{Y_2, Y_4, Y_6, ..., Y_{u-1}\}$.[4]

Based on the above formulations, the goal of an NCT model is to translate X_u to Y_u with certain types of dialogue history context.[5] Next, we will descibe the NCT model in our CSA-NCT training framework.

3.2 The NCT Model

The NCT model is based on transformer (Vaswani et al., 2017), which is composed of an encoder and a decoder as shown in Fig. 2.

Encoder. Following (Ma et al., 2020), the encoder takes $[\mathcal{C}_{X_u}; X_u]$ as input, where $[;]$ denotes the concatenation. In addition to the conventional embedding layer with only word embedding **WE** and position embedding **PE**, we additionally add a speaker embedding **SE** and a turn embedding **TE**. The final embedding $\mathbf{B}(x_i)$ of the input word x_i can be written as

$$\mathbf{B}(x_i) = \mathbf{WE}(x_i) + \mathbf{PE}(x_i) + \mathbf{SE}(x_i) + \mathbf{TE}(x_i),$$

where $\mathbf{WE} \in \mathbb{R}^{|V|\times d}$, $\mathbf{SE} \in \mathbb{R}^{2\times d}$ and $\mathbf{TE} \in \mathbb{R}^{|T|\times d}$.[6]

[4]For each item of $\{C_{X_u}, C_{Y_u}, C^{sx}_{Y_u}, C^{sy}_{Y_u}\}$, taking C_{X_u} for instance, we add the special token '[cls]' tag at the head of it and use another special token '[sep]' to delimit its included utterances, as in (Devlin et al., 2019).

[5]Here, we just take one translation direction (*i.e.*, En⇒Zh) as an example, which is similar for other directions.

[6]$|V|$, $|T|$ and d denote the size of shared vocabulary, maximum dialogue turns, and the hidden size, respectively.

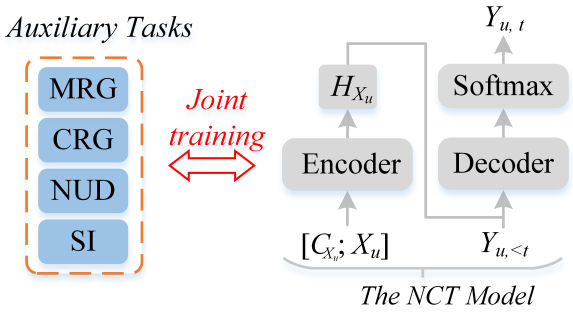

Figure 2: Architecture of the proposed CSA-NCT framework. The right part is the general NCT model, which is enhanced by four auxiliary tasks. The four auxiliary tasks including *monolingual response generation* (MRG), *cross-lingual response generation* (CRG), *next utterance discrimination* (NUD), and *speaker identification* (SI), are proposed to improve the coherence and speaker relevance of chat translation, which are presented in Fig. 3 in detail.

Then, the embedding is fed into the NCT encoder that has L identical layers, each of which is composed of a self-attention (SelfAtt) sub-layer and a feed-forward network (FFN) sub-layer.[7] Let $\mathbf{h}^l_e$ denote the hidden states of the l-th encoder layer, it is calculated as the following equations:

$$\mathbf{z}^l_e = \text{SelfAtt}(\mathbf{h}^{l-1}_e) + \mathbf{h}^{l-1}_e,$$
$$\mathbf{h}^l_e = \text{FFN}(\mathbf{z}^l_e) + \mathbf{z}^l_e,$$

where $\mathbf{h}^0_e$ is initialized as the embedding of input words. Particularly, words in $\mathcal{C}_{X_u}$ can only be attended to by those in X_u at the first encoder layer while $\mathcal{C}_{X_u}$ is masked at the other layers, which is the same implementation as in (Ma et al., 2020).

Decoder. The decoder also consists of L identical layers, each of which additionally includes a cross-attention (CrossAtt) sub-layer compared to the encoder. Let $\mathbf{h}^l_d$ denote the hidden states of the l-th decoder layer, it is computed as

$$\mathbf{z}^l_d = \text{SelfAtt}(\mathbf{h}^{l-1}_d) + \mathbf{h}^{l-1}_d,$$
$$\mathbf{c}^l_d = \text{CrossAtt}(\mathbf{z}^l_d, \mathbf{h}^L_e) + \mathbf{z}^l_d,$$
$$\mathbf{h}^l_d = \text{FFN}(\mathbf{c}^l_d) + \mathbf{c}^l_d,$$

where $\mathbf{h}^L_e$ is the top-layer encoder hidden states.

At each decoding time step t, $\mathbf{h}^L_{d,t}$ is fed into a linear transformation layer and a softmax layer to predict the probability distribution of the next target token:

$$p(Y_{u,t}|Y_{u,<t}, X_u, \mathcal{C}_{X_u}) = \text{Softmax}(\mathbf{W}_o\mathbf{h}^L_{d,t} + \mathbf{b}_o),$$

[7]The layer normalization is omitted for simplicity.

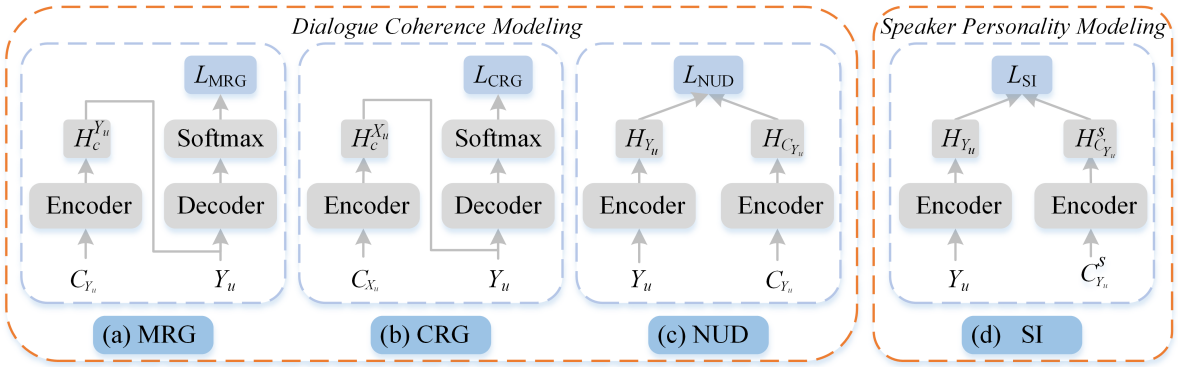

Figure 3: Overview of four auxiliary tasks. The encoder and the decoder of auxiliary tasks are shared with the NCT model. The encoder encodes not only source-side but also target-side history context to enhance its ability of representation.

where $Y_{u,<t}$ denotes the preceding tokens before the t-th time step in the utterance Y_u, $\mathbf{W}_o \in \mathbb{R}^{|V|\times d}$ and $\mathbf{b}_o \in \mathbb{R}^{|V|}$ are trainable parameters.

Finally, the training objective is as follows:

$$\mathcal{L}_{\text{NCT}} = -\sum_{t=1}^{|Y_u|} \log(p(Y_{u,t}|Y_{u,<t}, X_u, \mathcal{C}_{X_u})). \quad (3)$$

3.3 Auxiliary Tasks

We elaborately design four auxiliary tasks to incorporate the modeling of dialogue characteristics. The four auxiliary tasks are divided into two groups. The first group is for dialogue coherence modeling while the second is for speaker personality modeling. Together with the main chat translation task, the NCT model can be enhanced to generate more coherent and speaker-relevant translations through multi-task learning.

3.3.1 Dialogue Coherence Modeling

Many studies (Kuang et al., 2018; Wang et al., 2019b; Xiong et al., 2019; Wang and Wan, 2019; Huang et al., 2020) have indicated that the modeling of global textual coherence can lead to more coherent text generation. Inspired by this, we add two response generation tasks and an utterance discrimination task during the NCT model training. All the three tasks are related to the dialogue coherence of conversations, thus introducing the modeling of dialogue coherence into the NCT model.

Monolingual Response Generation (MRG). As illustrated in Fig. 3(a), given the dialogue history context $\mathcal{C}_{Y_u}$ in target language, the MRG task forces the NCT model to generate the corresponding utterance Y_u coherent to $\mathcal{C}_{Y_u}$. Particularly, we first use the encoder of the NCT model to encode $\mathcal{C}_{Y_u}$, and then use the NCT decoder to predict Y_u. The training objective of this task can be formulated as:

$$\mathcal{L}_{\text{MRG}} = -\sum_{t=1}^{|Y_u|} \log(p(Y_{u,t}|\mathcal{C}_{Y_u}, Y_{u,<t})),$$
$$p(Y_{u,t}|\mathcal{C}_{Y_u}, Y_{u,<t}) = \text{Softmax}(\mathbf{W}_m \mathbf{h}^L_{d,t} + \mathbf{b}_m),$$

where $\mathbf{h}^L_{d,t}$ is the top-layer decoder hidden state at the t-th decoding step, $\mathbf{W}_m$ and $\mathbf{b}_m$ are trainable parameters.

Cross-lingual Response Generation (CRG). The CRG task is similar to the MRG as shown in Fig. 3(b), where the NCT model is trained to generate the corresponding utterance Y_u in target language which is coherent to the given dialogue history context $\mathcal{C}_{X_u}$ in source language. We first use the encoder of the NCT model to encode $\mathcal{C}_{X_u}$, and then use the NCT decoder to predict Y_u. The training objective of this task can be formulated as:

$$\mathcal{L}_{\text{CRG}} = -\sum_{t=1}^{|Y_u|} \log(p(Y_{u,t}|\mathcal{C}_{X_u}, Y_{u,<t})),$$
$$p(Y_{u,t}|\mathcal{C}_{X_u}, Y_{u,<t}) = \text{Softmax}(\mathbf{W}_c \mathbf{h}^L_{d,t} + \mathbf{b}_c),$$

where $\mathbf{h}^L_{d,t}$ denotes the top-layer decoder hidden state at the t-th decoding step, $\mathbf{W}_{crg}$ and $\mathbf{b}_{crg}$ are trainable parameters.

Note that in the above two response generation tasks, we use the same set of NCT model parameters except for the softmax layer (*i.e.*, $\mathbf{W}_m$, $\mathbf{b}_m$, $\mathbf{W}_c$ and $\mathbf{b}_c$).

Next Utterance Discrimination (NUD). As shown in Fig. 3(c), we design the NUD task to

distinguish whether the translated text is coherent to be the next utterance of the given dialogue history context. Concretely, we construct positive and negative samples of context-utterance pairs from the chat corpus. A positive sample ($\mathcal{C}_{Y_u}$, Y_{u^+}) with the label $\ell = 1$ consists of the target utterance Y_u and its dialogue history context $\mathcal{C}_{Y_u}$. A negative sample ($\mathcal{C}_{Y_u}$, Y_{u^-}) with the label $\ell = 0$ consists of the identical $\mathcal{C}_{Y_u}$ and a randomly selected utterance Y_{u^-} from the training set. Formally, the training objective of NUD is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{NUD}} = &-\log(p(\ell=1|\mathcal{C}_{Y_u}, Y_{u^+})) \\ &-\log(p(\ell=0|\mathcal{C}_{Y_u}, Y_{u^-})).\end{aligned} \tag{4}$$

For a training sample ($\mathcal{C}_{Y_u}$, Y_u), to estimate the probability in Eq. 4 for discrimination, we first obtain the representations $\mathbf{H}_{Y_u}$ of the target utterance Y_u and $\mathbf{H}_{\mathcal{C}_{Y_u}}$ of the given dialogue history context $\mathcal{C}_{Y_u}$ using the NCT encoder. Specifically, $\mathbf{H}_{Y_u}$ is calculated as $\frac{1}{|Y_u|}\sum_{t=1}^{|Y_u|}\mathbf{h}_{e,t}^{L}$ while $\mathbf{H}_{\mathcal{C}_{Y_u}}$ is defined as the encoder hidden state $\mathbf{h}_{e,0}^{L}$ of the prepended special token '[cls]' of $\mathcal{C}_{Y_u}$. Then, the concatenation of $\mathbf{H}_{Y_u}$ and $\mathbf{H}_{\mathcal{C}_{Y_u}}$ is fed into a binary NUD classifier, which is an extra fully-connected layer on top of the NCT encoder:

$$p(\ell=1|\mathcal{C}_{Y_u}, Y_u) = \text{Softmax}(\mathbf{W}_n[\mathbf{H}_{Y_u}; \mathbf{H}_{\mathcal{C}_{Y_u}}]),$$

where $\mathbf{W}_n$ is the trainable parameter of the NUD classifier and the bias term is omitted for simplicity.

3.3.2 Speaker Personality Modeling

A dialogue always involves speakers who have different personalities, which is a salient characteristic of conversations. Therefore, we design a speaker identification task that incorporates the modeling of speaker personality into the NCT model, making the translated utterance more speaker-relevant.

Speaker Identification (SI). As explored in (Bak and Oh, 2019; Wu et al., 2020; Liang et al., 2021b; Lin et al., 2021), the history utterances of a speaker can reflect a distinctive personality. Fig. 3(d) depicts the SI task in detail, where the NCT model is used to distinguish whether a translated utterance and a given speaker-specific history utterances are spoken by the same speaker. We also construct positive and negative training samples from the chat corpus. A positive sample ($\mathcal{C}_{Y_u}^{sx}$, Y_u) with the label $\ell = 1$ consists of the target utterance Y_u and the speaker sx-specific history context $\mathcal{C}_{Y_u}^{sx}$, because Y_u is the translation of the utterance originally spoken by the speaker sx. A negative sample

Algorithm 1: Optimization Algorithm

Input: Sentence-level/Chat-level translation data $\mathcal{D}^s$/ $\mathcal{D}^c$, Sentence-level/Chat-level MaxStep T_1/T_2, CoherenceMaxStep T_2, SpeakerMaxStep T_2

Init: θ

$t_1 = 0$ (*Training sentence-level NMT model*)

for $t_1 < T_1$ **do**

  Randomly sample a batch k from $\mathcal{D}^s$.

  Compute $\mathcal{L}_{\text{S-NMT}}$.

  Update the parameters of the standard transformer model using Adam.

Output: θ

Init: Θ using θ, $\alpha = 1.0$, $\beta = 1.0$

$t_2 = 0$ (*Training chat-level NMT model*)

for $t_2 < T_2$ **do**

  Randomly sample a batch k from $\mathcal{D}^c$.

  Compute $\mathcal{L}_{\text{MRG}}$, $\mathcal{L}_{\text{CRG}}$, $\mathcal{L}_{\text{NUD}}$, $\mathcal{L}_{\text{SI}}$, and $\mathcal{L}_{\text{NCT}}$.

  Update the parameters of the CSA-NCT model with respect to $\mathcal{J}$ using Adam.

  $d_1 = \alpha * t_2/T_2, \quad d_2 = \beta * t_2/T_2$

  $\alpha = max(0, \alpha - d_1)$

  $\beta = max(0, \beta - d_2)$

Output: Θ

($\mathcal{C}_{Y_u}^{sy}$, Y_u) with the label $\ell = 0$ consists of the target utterance Y_u and the speaker sy-specific history context $\mathcal{C}_{Y_u}^{sy}$. Formally, the training objective of SI is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{SI}} = &-\log(p(\ell=1|\mathcal{C}_{Y_u}^{sx}, Y_u)) \\ &-\log(p(\ell=0|\mathcal{C}_{Y_u}^{sy}, Y_u)).\end{aligned} \tag{5}$$

For a training sample ($\mathcal{C}_{Y_u}^{s}$, Y_u) with $s \in \{sx, sy\}$, we also use the NCT encoder to obtain the representations $\mathbf{H}_{Y_u}$ of the target utterance Y_u and $\mathbf{H}_{\mathcal{C}_{Y_u}^{s}}$ of the given speaker-specific history context $\mathcal{C}_{Y_u}^{s}$. Similar to the NUD task, $\mathbf{H}_{Y_u} = \frac{1}{|Y_u|}\sum_{t=1}^{|Y_u|}\mathbf{h}_{e,t}^{L}$ and the $\mathbf{h}_{e,0}^{L}$ of $\mathcal{C}_{Y_u}^{s}$ is used as $\mathbf{H}_{\mathcal{C}_{Y_u}^{s}}$. Then, to estimate the probability in Eq. 5, the concatenation of $\mathbf{H}_{Y_u}$ and $\mathbf{H}_{\mathcal{C}_{Y_u}^{s}}$ is fed into a binary SI classifier, which is another fully-connected layer on top of the NCT encoder:

$$p(\ell=1|\mathcal{C}_{Y_u}^{s}, Y_u) = \text{Softmax}(\mathbf{W}_s[\mathbf{H}_{Y_u}; \mathbf{H}_{\mathcal{C}_{Y_u}^{s}}]),$$

where $\mathbf{W}_s$ is the trainable parameter of the SI classifier and the bias term is also omitted.

3.4 Training and Inference

For training, with the main chat translation task and four auxiliary tasks, the total training objective is finally formulated as

$$\mathcal{J} = \mathcal{L}_{\text{NCT}} + \alpha(\mathcal{L}_{\text{MRG}} + \mathcal{L}_{\text{CRG}} + \mathcal{L}_{\text{NUD}}) + \beta\mathcal{L}_{\text{SI}}, \quad (6)$$

where α and β are balancing hyper-parameters for the trade-off between $\mathcal{L}_{\text{NCT}}$ and the other auxiliary objectives. Algorithm 1 summarizes the training procedure of the above multi-task learning process, where θ refers to the parameters of our NCT model and Θ refers to the whole set of parameters including both θ and the parameters of the additional classifiers for auxiliary tasks.

During inference, the four auxiliary tasks are not involved and only the NCT model (θ) is used to conduct chat translation.

4 Experiments

4.1 Datasets and Metrics

Datasets. As shown in Algorithm 1, the training of our CSA-NCT framework consists of two stages: (1) pre-train the model on a large-scale sentence-level NMT corpus (WMT20[8]); (2) fine-tune on the chat translation corpus (BConTrasT (Farajian et al., 2020) and BMELD (Liang et al., 2021a)). The dataset details (*e.g.*, splits of training, validation or test sets) are described in Appendix A.

Metrics. For fair comparison, we use the SacreBLEU[9] (Post, 2018) and TER (Snover et al., 2006) 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 Zh⇒En, we report case-insensitive score. For En⇒Zh, the reported SacreBLEU is at the character level.

4.2 Implementation Details

In this paper, we adopt the settings of standard *Transformer-Base* and *Transformer-Big* in (Vaswani et al., 2017) and follow the main setting in (Liang et al., 2021a). Specifically, in *Transformer-Base*, we use 512 as hidden size (*i.e.*, d), 2048 as filter size and 8 heads in multihead attention. In *Transformer-Big*, we use 1024 as hidden size, 4096 as filter size, and 16 heads in multihead attention. All our Transformer models contain L = 6 encoder layers and L = 6 decoder layers and all models are trained using THUMT (Tan et al., 2020) framework. The training step for the first pre-training stage is set to T_1 = 200,000 while that of the second fine-tuning stage is set to T_2 = 5,000. The batch size for each GPU is set to 4096 tokens. All experiments in the first stage are conducted utilizing 8 NVIDIA Tesla V100 GPUs, while we use 4 GPUs for the second stage, *i.e.*, fine-tuning. That gives us about 8*4096 and 4*4096 tokens per update for all experiments in the first-stage and second-stage, respectively. All models are optimized using Adam (Kingma and Ba, 2014) with β_1 = 0.9 and β_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. $|T|$ is set to 10. Following (Liang et al., 2021a), we set the number of preceding sentences to 3 in all experiments. The criterion for selecting hyper-parameters is the BLEU score on validation sets for both tasks. During inference, the beam size is set to 4, and the length penalty is 0.6 among all experiments.

	Setting	En⇒De	En⇒Zh
Big	Fixed α and β	60.91/24.6	29.69/55.4
	Dynamic α and β	61.27/24.3	30.52/54.6

Table 1: The BLEU/TER score (%) results on the validation sets.

4.3 Effect of α and β

We also investigate the effect of balancing factor α and β, where α and β gradually decrease from 1 to 0 over 5,000 steps, which is similar to (Zhao et al., 2020). "Fixed α and β" means we keep $\alpha = \beta = 1$ across the training. "Dynamic α and β" denotes decaying α and β with the training step of auxiliary tasks. The results of Tab. 1 show that "Dynamic α and β" gives better performance than "Fixed α and β". Therefore, we apply this dynamic strategy in the following experiments.

4.4 Comparison Models

Baseline Sentence-Level NMT Models.

- **Transformer** (Vaswani et al., 2017): The de-facto NMT model trained on sentence-level NMT corpus.
- **Transformer+FT** (Vaswani et al., 2017): The NMT model that is directly fine-tuned on the

[8] http://www.statmt.org/wmt20/translation-task.html

[9] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.13

	Models	En⇒De		De⇒En		En⇒Zh		Zh⇒En	
		BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓
Sentence-Level 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	59.57	26.2	25.22	62.8	21.59	56.7
Context-Aware NMT models (Base)	Dia-Transformer+FT	58.33	26.8	59.09	26.2	24.96	63.7	20.49	60.1
	Doc-Transformer+FT	58.15	27.1	59.46	25.7	24.76	63.4	20.61	59.8
	Gate-Transformer+FT	58.48	26.6	59.53	26.1	25.34	62.5	21.03	56.9
	CSA-NCT (Ours)	**59.50**††	**25.7**††	**60.65**††	**25.4**†	**27.77**††	**60.0**††	**22.36**†	**55.9**†
Sentence-Level NMT models (Big)	Transformer	40.53	42.2	49.90	33.3	22.81	69.6	19.58	57.7
	Transformer+FT	59.01	26.0	59.98	25.9	26.95	60.7	22.15	56.1
Context-Aware NMT models (Big)	Dia-Transformer+FT	58.68	26.8	59.63	26.0	26.72	62.4	21.09	58.1
	Doc-Transformer+FT	58.61	26.5	59.98	25.4	26.45	62.6	21.38	57.7
	Gate-Transformer+FT	58.94	26.2	60.08	25.5	27.13	60.3	22.26	55.8
	CSA-NCT (Ours)	**60.64**††	**25.3**†	**61.21**††	**24.9**†	**28.86**††	**58.7**††	**23.69**††	**54.7**††

Table 2: Results on the test sets of BConTrasT (En⇔De) and BMELD (En⇔Zh) 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. All "+FT" models apply the same two-stage training strategy with our CSA-NCT model for fair comparison.

chat translation data after being pre-trained on sentence-level NMT corpus.

Existing Context-Aware NMT Systems.

- **Dia-Transformer+FT** (Maruf et al., 2018): The original model is RNN-based and an additional encoder is used to incorporate the mixed-language dialogue history. We re-implement it based on Transformer where an additional encoder layer is used to introduce the dialogue history into NMT model.
- **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 dialogue history.
- **Gate-Transformer+FT** (Zhang et al., 2018): A document-aware Transformer that uses a gate to incorporate the context information. Note that we share the Transformer encoder to obtain the context representation instead of utilizing the additional context encoder, which performs better in our experiments.

4.5 Main Results

In Tab. 2, We report the main results on En⇔De and En⇔Zh under *Base* and *Big* settings. For comparison, as in § 4.4, "Transformer" and "Transformer+FT" are sentence-level baselines while "Dia-Transformer+FT", "Doc-Transformer+FT" and "Gate-Transformer+FT" are the existing context-aware NMT systems re-implemented by us. Particularly, "CSA-NCT" represents our proposed approach.

Results on En⇔De. Under the *Base* setting, our model substantially outperforms the sentence-level/context-aware baselines by a large margin (*e.g.*, the previous best "Gate-Transformer+FT"), 1.02↑ on En⇒De and 1.12↑ on De⇒En. In term of TER, CSA-NCT also performs better on the two directions, 0.9↓ and 0.7↓ lower than "Gate-Transformer+FT" (the lower the better), respectively. Under the *Big* setting, on En⇒De and De⇒En, our model consistently surpasses the baselines and other existing systems again.

Results on En⇔Zh. We also conduct experiments on the BMELD dataset. Concretely, on En⇒Zh and Zh⇒En, our model also presents notable improvements over all comparison models by at least 2.43↑ and 0.77↑ BLEU gains under the *Base* setting, and by 1.73↑ and 1.43↑ BLEU gains under the *Big* setting, respectively. These results demonstrate the effectiveness and generalizability of our model across different language pairs.

5 Analysis

5.1 Ablation Study

Effect of Each Auxiliary Task Group. We conduct ablation studies to investigate the effects of the two groups (DCM and SPM) of auxiliary tasks. The results under the *Big* setting are listed in Tab. 3.

#	Models	En⇒De BLEU↑	En⇒De TER↓	De⇒En BLEU↑	De⇒En TER↓
0	Baseline	60.40	25.0	61.68	24.9
1	*w/* DCM	61.05†† (+0.65)	24.4††	62.63†† (+0.95)	24.5†
2	*w/* SPM	60.57 (+0.17)	24.8	61.97 (+0.29)	24.7

Table 3: Ablation results on the validation sets of each auxiliary task group under the *Big* setting. "Baseline" represents the NCT model without any auxiliary task. "DCM": dialogue coherence modeling, including MRG, CRG, NUD. "SPM": speaker personality modeling, *i.e.*, SI. "†" and "††" indicate the improvement over the result of the baseline model is statistically significant with $p < 0.05$ and $p < 0.01$), respectively.

#	Models	En⇒De BLEU↑	En⇒De TER↓	De⇒En BLEU↑	De⇒En TER↓
0	Baseline	60.40	25.0	61.68	24.9
1	*w/* MRG	61.00†† (+0.60)	24.4††	62.37†† (+0.69)	24.5†
2	*w/* CRG	60.68 (+0.28)	24.6†	62.14† (+0.46)	24.8
3	*w/* NUD	60.82† (+0.42)	24.7	62.32†† (+0.64)	24.7
4	*w/* SI	60.57 (+0.17)	24.8	61.97 (+0.29)	24.7

Table 4: Ablation results on the validation sets of each auxiliary task under the *Big* setting. "†" and "††" indicate the improvement over the result of the baseline model is statistically significant with $p < 0.05$ and $p < 0.01$, respectively.

We have the following findings: (1) DCM substantially improves the NCT model in terms of both BLEU and TER metrics, which demonstrates modeling coherence is beneficial for better translations. (2) SPM makes slight contributions to the NCT model in terms of BLEU, which is less significant than DCM. However, further human evaluation in § 5.3 will show that our model can keep the personality consistent with the original speaker.

Effect of Each Auxiliary Task. We also investigate the effect of each auxiliary task by adding a single task at a time. In Tab. 4, rows 1∼4 denote singly adding on the corresponding auxiliary task with the main chat translation task, each of which shows a positive impact on the model performance (rows 1∼4 vs. row 0).

5.2 Dialogue Coherence

Following (Lapata and Barzilay, 2005; Xiong et al., 2019), we measure dialogue coherence as sentence similarity, which is determined by the cosine similarity between two sentences s_1 and s_2:

$$sim(s_1, s_2) = \cos(f(s_1), f(s_2)),$$

Models	1-th Pr.	2-th Pr.	3-th Pr.
Transformer	65.02	60.37	56.59
Transformer+FT	65.87	61.04	57.14
Dia-Transformer+FT	65.53	60.84	57.09
Doc-Transformer+FT	65.69	60.93	57.13
Gate-Transformer+FT	65.96	61.35	57.45
CSA-NCT (Ours)	66.57††	61.78††	57.83††
Human Reference	**66.63**	**61.90**	**57.95**

Table 5: Results (%) of dialogue coherence in terms of sentence similarity on the test set of BConTrasT in De⇒En direction under the *Base* setting. The "**#-th Pr.**" denotes the #-th preceding utterance to the current one. "††" indicates the improvement over the best result of all other comparison models is statistically significant ($p < 0.01$).

Models	Coh.	Spe.	Flu.
Transformer	0.540	0.485	0.590
Transformer+FT	0.590	0.530	0.635
Dia-Transformer+FT	0.580	0.525	0.625
Doc-Transformer+FT	0.595	0.525	0.630
Gate-Transformer+FT	0.605	0.540	0.635
CSA-NCT (Ours)	**0.635**	**0.575**	**0.655**

Table 6: Results of Human evaluation (Zh⇒En, *Base*). "**Coh.**": Coherence. "**Spe.**": Speaker. "**Flu.**": Fluency.

where $f(s_i) = \frac{1}{|s_i|}\sum_{\mathbf{w}\in s_i}(\mathbf{w})$ and $\mathbf{w}$ is the vector for word w. We use Word2Vec[10] (Mikolov et al., 2013) trained on a dialogue dataset[11] to obtain the distributed word vectors whose dimension is set to 100.

Tab. 5 shows the measured coherence of different models on the test set of BConTrasT in De⇒En direction. It shows that our CSA-NCT produces more coherent translations compared to baselines and other existing systems (significance test, $p < 0.01$).

5.3 Human Evaluation

Inspired by (Bao et al., 2020; Farajian et al., 2020), we use three criteria for human evaluation: (1) **Coherence** measures whether the translation is semantically coherent with the dialogue history; (2) **Speaker** measures whether the translation preserves the personality of the speaker; (3) **Fluency** measures whether the translation is fluent and gram-

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

[11]Due to no available German dialogue datasets, we choose Taskmaster-1 (Byrne et al., 2019), where the English side of BConTrasT (Farajian et al., 2020) also comes from it.

matically correct.

First, we randomly sample 200 conversations from the test set of BMELD in Zh⇒En direction. Then, we use the 6 models in Tab. 6 to generate the translated utterances of these sampled conversations. Finally, we assign the translated utterances and their corresponding dialogue history utterances in target language to three postgraduate human annotators, and ask them to make evaluations from the above three criteria.

The results in Tab. 6 show that our model generates more coherent, speaker-relevant, and fluent translations compared with other models (significance test, $p < 0.05$), indicating the superiority of our model. The inter-annotator agreements calculated by the Fleiss' kappa (Fleiss and Cohen, 1973) are 0.506, 0.548, and 0.497 for coherence, speaker and fluency, respectively, indicating "Moderate Agreement" for all four criteria. We also present one case study in Appendix B.

6 Related Work

Chat NMT. Little prior work is available due to the lack of human-annotated publicly available data (Farajian et al., 2020). Therefore, some existing studies (Wang et al., 2016; Maruf et al., 2018; Zhang and Zhou, 2019; Rikters et al., 2020) mainly pay attention to designing methods to automatically construct the subtitle corpus, which may contain noisy bilingual utterances. Recently, Farajian et al. (2020) organize the WMT20 chat translation task and first provide a chat corpus post-edited by humans. More recently, based on document-level parallel corpus, Wang et al. (2021) propose to jointly identify omissions and typos within dialogue along with translating utterances by using the context. As a concurrent work, Liang et al. (2021a) provide a clean bilingual dialogue dataset and design a variational framework for NCT. Different from them, we focus on introducing the modeling of dialogue coherence and speaker personality into the NCT model with multi-task learning to promote the translation quality.

Context-Aware NMT. In a sense, chat MT can be viewed as a special case of context-aware MT that has many related studies (Gong et al., 2011; Jean et al., 2017; Wang et al., 2017b; Zheng et al., 2020; Yang et al., 2019; Kang et al., 2020; Li et al., 2020; Chen et al., 2020; Ma et al., 2020). Typically, they resort to extending conventional NMT models for exploiting the context. Although these models can be directly applied to the chat translation scenario, they cannot explicitly capture the inherent dialogue characteristics and usually lead to incoherent and speaker-irrelevant translations.

7 Conclusion

In this paper, we propose to enhance the NCT model by introducing the modeling of the inherent dialogue characteristics, *i.e.*, dialogue coherence and speaker personality. We train the NCT model with the four well-designed auxiliary tasks, *i.e.*, MRG, CRG, NUD and SI. Experiments on En⇔De and En⇔Zh 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 more coherent and speaker-relevant translations.

Acknowledgements

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.

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.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Mod-

eling discourse structure for document-level neural machine translation. *CoRR*.

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*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of EMNLP*, pages 9230–9240, Online.

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. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021a. Modeling bilingual conversational characteristics for neural chat translation. In *Proceedings of ACL*, pages 5711–5724, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021b. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. *Proceedings of AAAI*, pages 13343–13352.

Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. In *Proceedings of ACL/IJCNLP*, pages 4008–4018. Association for Computational Linguistics.

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.

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.

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*.

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.

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. 2019a. 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.

Weichao Wang, Shi Feng, Daling Wang, and Yifei Zhang. 2019b. Answer-guided and semantic coherent question generation in open-domain conversation. In *Proceedings of EMNLP-IJCNLP*, pages 5066–5076, Hong Kong, China.

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.

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.

Yufan Zhao, Can Xu, and Wei Wu. 2020. Learning a simple and effective model for multi-turn response generation with auxiliary tasks. In *Proceedings of EMNLP*, pages 3472–3483, Online. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of IJCAI*.

A Datasets

As mentioned in § 4.1, our experiments involve the dataset WMT20 for pre-training and two chat translation corpus, BConTrasT (Farajian et al., 2020) and BMELD (Liang et al., 2021a). The statistics about the splits of training, validation, and test sets are shown in Tab. 7.

WMT20. Following (Liang et al., 2021a), for En⇔De, we combine six corpora including Europal, ParaCrawl, CommonCrawl, TildeRapid, News-Commentary, and WikiMatrix. For En⇔Zh, we combine News Commentary v15, Wiki Titles v2, UN Parallel Corpus V1.0, CCMT Corpus, and WikiMatrix. First, we filter out duplicate sentence pairs and remove those 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, we apply BPE (Sennrich et al., 2016) with 32K merge operations to obtain subwords. Finally, we obtain 45,541,367 sentence pairs for En⇔De and 22,244,006 sentence pairs for En⇔Zh, respectively.

Datasets	# 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⇒Zh	1,036	108	274	5,560	567	1,466
Zh⇒En	1,036	108	274	4,427	517	1,135

Table 7: Statistics of chat translation data.

Models	**En⇒De**	**De⇒En**	**En⇒Zh**	**Zh⇒En**
Transformer (Base)	39.88	40.72	32.55	24.42
Transformer (Big)	41.35	41.56	33.85	24.86

Table 8: The BLEU scores on the *newstest2019* of the first stage.

We test the model performance of the first stage on *newstest2019*. The results are shown in Tab. 8.

BConTrasT. The dataset[12] 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[13] who are native German speakers. Having the conversations in both languages allows us to simulate bilingual conversations in which one speaker (customer), speaks in German and the other speaker (agent), responds in English.

BMELD. The dataset is a recently released English⇔Chinese bilingual dialogue dataset, provided by Liang et al. (2021a). Based on the dialogue dataset in the MELD (originally in English) (Poria et al., 2019)[14], they firstly crawled the corresponding Chinese translations from `https://www.zimutiantang.com/` and then manually post-edited them according to the dialogue history by native Chinese speakers who are post-graduate students majoring in English. Finally, following (Farajian et al., 2020), they assume 50% speakers as Chinese speakers to keep data balance for Zh⇒En translations and build the bilingual MELD (BMELD). For the Chinese, we follow them to segment the sentence using Stanford

[12] https://github.com/Unbabel/BConTrasT

[13] www.unbabel.com

[14] 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.

Figure 4: An illustrative case of bilingual conversation.

CoreNLP toolkit[15].

B Case Study

In this section, we deliver an illustrative case in Fig. 4 to show different outputs among the comparison models and ours.

Dialogue Coherence and Speaker Personality. For the case in Fig. 4, we find that all comparison models cannot generate coherent translated utterences. The reason may be that they fail to capture contextual clues, *i.e.*, "*boat*". By contrast, we explicitly introduce the modeling of preceding context through auxiliary tasks and thus obtain satisfactory results. Meanwhile, we observe that the sentence-level models and the context-aware models cannot preserve the speaker personality information, *e.g.*, *joy* emotion, even though context-aware models incorporate the bilingual conversational history into the encoder.

The case shows that our CSA-NCT model enhanced by the four auxiliary tasks yields coherent and speaker-relevant translations, demonstrating its effectiveness and superiority.

[15]https://stanfordnlp.github.io/CoreNLP/index.html