Scheduled Multi-task Learning for Neural Chat Translation

Yunlong Liang[1]*, Fandong Meng[2], Jinan Xu[1]†, Yufeng Chen[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,jaxu,chenyf}@bjtu.edu.cn
{fandongmeng,withtomzhou}@tencent.com

Abstract

Neural Chat Translation (NCT) aims to translate conversational text into different languages. Existing methods mainly focus on modeling the bilingual dialogue characteristics (*e.g.*, coherence) to improve chat translation via multi-task learning on small-scale chat translation data. Although the NCT models have achieved impressive success, it is still far from satisfactory due to insufficient chat translation data and simple joint training manners. To address the above issues, we propose a scheduled multi-task learning framework for NCT. Specifically, we devise a three-stage training framework to incorporate the large-scale in-domain chat translation data into training by adding a second pre-training stage between the original pre-training and fine-tuning stages. Further, we investigate where and how to schedule the dialogue-related auxiliary tasks in multiple training stages to effectively enhance the main chat translation task. Extensive experiments on four language directions (English↔Chinese and English↔German) verify the effectiveness and superiority of the proposed approach. Additionally, we will make the large-scale in-domain paired bilingual dialogue dataset publicly available for the research community.[1]

1 Introduction

A cross-lingual conversation involves speakers in different languages (*e.g.*, one speaking in Chinese and another in English), where a chat translator can be applied to help them communicate in their native languages. The chat translator bilaterally converts the language of bilingual conversational text, *e.g.* from Chinese to English and vice versa (Wang et al., 2016a; Farajian et al., 2020; Liang et al., 2021a, 2022).

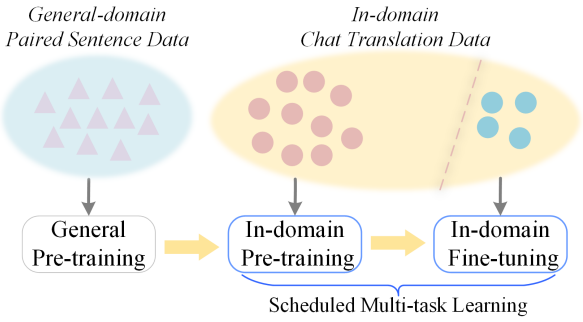


Figure 1: The overall three-stage training framework.

Generally, since the bilingual dialogue corpus is scarce, researchers (Bao et al., 2020; Wang et al., 2020; Liang et al., 2021a,d) resort to making use of the large-scale general-domain data through the pre-training-then-fine-tuning paradigm as done in many context-aware neural machine translation models (Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Miculicich et al., 2018; Tu et al., 2018; Voita et al., 2018, 2019a,b; Yang et al., 2019; Wang et al., 2019; Maruf et al., 2019; Ma et al., 2020, etc), having made significant progress. However, conventional pre-training on large-scale general-domain data usually learns general language patterns, which is also aimless for capturing the useful dialogue context to chat translation, and fine-tuning usually suffers from insufficient supervised data (about 10k bilingual dialogues). Some studies (Gu et al., 2020; Gururangan et al., 2020; Liu et al., 2021; Moghe et al., 2020; Wang et al., 2020; Ruder, 2021) have shown that learning domain-specific patterns by additional pre-training is beneficial to the models. To this end, we firstly construct the large-scale in-domain chat translation data[2]. And to

*Work was done when Yunlong was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.

†Jinan Xu is the corresponding author.

[1]The code and in-domain data are publicly available at: https://github.com/XL2248/SML

[2]Firstly, to build the data, for English↔Chinese (En↔Zh), we crawl two consecutive English and Chinese movie subtitles (not aligned). For English↔German (En↔De), we download two consecutive English and German movie subtitles (not aligned). Then, we use several advanced technologies to align En↔Zh and En↔De subtitles. Finally, we obtain the paired bilingual dialogue dataset. Please refer to § 3.1 for details.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4375 - 4388

incorporate it for learning domain-specific patterns, we then propose a three-stage training framework via adding a second pre-training stage between general pre-training and fine-tuning, as shown in Fig. 1.

To further improve the chat translation performance through modeling dialogue characteristics (*e.g.*, coherence), inspired by previous studies (Phang et al., 2020; Liang et al., 2021d; Pruksachatkun et al., 2020), we incorporate several dialogue-related auxiliary tasks to our three-stage training framework. Unfortunately, we find that simply introducing all auxiliary tasks in the conventional multi-task learning manner does not obtain significant cumulative benefits as we expect. It indicates that the simple joint training manner may limit the potential of these auxiliary tasks, which inspires us to investigate where and how to make these auxiliary tasks work better for the main NCT task.

To address the above issues, we present a **S**cheduled **M**ulti-task **L**earning framework (SML) for NCT, as shown in Fig. 1. Firstly, we propose a three-stage training framework to introduce our constructed in-domain chat translation data for learning domain-specific patterns. Secondly, to make the most of auxiliary tasks for the main NCT task, **where**: we analyze in which stage these auxiliary tasks work well and find that they are *different strokes for different folks*. Therefore, to fully exert their advantages for enhancing the main NCT task, **how**: we design a gradient-based strategy to dynamically schedule them at each training step in the last two training stages, which can be seen as a fine-grained joint training manner. In this way, the NCT model is effectively enhanced to capture both domain-specific patterns and dialogue-related characteristics (*e.g.*, coherence) in conversation, which thus can generate better translation results.

We validate our SML framework on two datasets: BMELD (Liang et al., 2021a) (En↔Zh) and BConTrasT (Farajian et al., 2020) (En↔De). Experimental results show that our model gains consistent improvements on four translation tasks in terms of both BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores, demonstrating its effectiveness and generalizability. Human evaluation further suggests that our model can produce more coherent and fluent translations compared to the previous related methods.

Our contributions are summarized as follows:

- We propose a scheduled multi-task learning framework with three training stages, where a gradient-based scheduling strategy is designed to fully exert the auxiliary tasks' advantages for the main NCT task, for higher translation quality.
- Extensive experiments on four chat translation tasks show that our model achieves new state-of-the-art performance and outperforms the existing NCT models by a significant margin.
- We contribute two large-scale in-domain paired bilingual dialogue corpora (28M for En↔Zh and 18M for En↔De) to the research community.

2 Background: Conventional Multi-task Learning for NCT

We introduce the conventional multi-task learning framework (Liang et al., 2021d) for NCT, which includes four parts: *problem formalization* (§ 2.1), *the NCT model* (§ 2.2), *existing three auxiliary tasks* (§ 2.3), and *training objective* (§ 2.4).

2.1 Problem Formalization

In a bilingual conversation, we assume the two speakers have alternately given utterances in different languages for u turns, resulting in $X_1, X_2, X_3, ..., X_u$ and $Y_1, Y_2, Y_3, ..., Y_u$ on the source and target sides, respectively. Among these utterances, $X_1, X_3, X_5, ..., X_u$ are originally spoken and $Y_1, Y_3, Y_5, ..., Y_u$ are the corresponding translations in the target language. Similarly, $Y_2, Y_4, Y_6, ..., Y_{u-1}$ are originally spoken and $X_2, X_4, X_6, ..., X_{u-1}$ are the translated utterances in the source language. According to languages, we define the dialogue history context of X_u on the source side as $\mathcal{C}_{X_u}$=$\{X_1, X_2, X_3, ..., X_{u-1}\}$ and that of Y_u on the target side as C_{Y_u}=$\{Y_1, Y_2, Y_3, ..., Y_{u-1}\}$.[3]

The goal of an NCT model is to translate X_u to Y_u with dialogue history context $\mathcal{C}_{X_u}$ and C_{Y_u}.

2.2 The NCT Model

The NCT model (Ma et al., 2020; Liang et al., 2021d) utilizes the standard transformer (Vaswani et al., 2017) architecture with an encoder and a decoder[4].

[3]For each of $\{C_{X_u}, C_{Y_u}\}$, we add the special token '[CLS]' tag at the head of it and use another token '[SEP]' to delimit its included utterances, as in Devlin et al. (2019).

[4]Here, we just describe some adaptions to the NCT model, and please refer to Vaswani et al. (2017) for more details.

In the encoder, it takes $[\mathcal{C}_{X_u}; X_u]$ as input, where $[;]$ denotes the concatenation. The input embedding consists of word embedding $\mathbf{WE}$, position embedding $\mathbf{PE}$, and turn embedding $\mathbf{TE}$:

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

where $\mathbf{WE} \in \mathbb{R}^{|V| \times d}$ and $\mathbf{TE} \in \mathbb{R}^{|T| \times d}$.[5] When computation in the encoder, words in $\mathcal{C}_{X_u}$ can only be attended 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).

In the decoder, at each decoding time step t, the top-layer (L-th) decoder hidden state $\mathbf{h}_{d,t}^{L}$ is fed into a softmax layer to predict the probability distribution of the next target token:

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

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 loss is defined as follows:

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

2.3 Existing Auxiliary Tasks

To generate coherent translation, Liang et al. (2021d) present Monolingual Response Generation (MRG) task, Cross-lingual Response Generation (XRG) task, and Next Utterance Discrimination (NUD) task during the NCT model training.

MRG. Given the dialogue context $\mathcal{C}_{Y_u}$ in the target language, it forces the NCT model to generate the corresponding utterance Y_u coherent to $\mathcal{C}_{Y_u}$. Particularly, the encoder of the NCT model is used to encode $\mathcal{C}_{Y_u}$, and the NCT decoder predicts Y_u. The training objective of this task is formulated as:

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

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

XRG. Similar to MRG, the NCT model is also jointly trained to generate the corresponding utterance Y_u which is coherent to the given dialogue history context $\mathcal{C}_{X_u}$ in the source language:

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

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are trainable parameters.

NUD. The NUD task aims to distinguish whether the translated text is coherent to be the next utterance of the given dialogue history context. Specifically, the positive and negative samples are firstly constructed: (1) the 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 $\mathcal{C}_{Y_u}$; (2) the 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 preceding context of Y_u. Formally, the training objective of NUD is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{NUD}} = &-\log(p(\ell = 1|\mathcal{C}_{Y_u}, Y_{u^+})) \\ &-\log(p(\ell = 0|\mathcal{C}_{Y_u}, Y_{u^-})),\end{aligned}$$
$$p(\ell = 1|\mathcal{C}_{Y_u}, Y_u) = \text{Softmax}(\mathbf{W}_n[\mathbf{H}_{Y_u}; \mathbf{H}_{\mathcal{C}_{Y_u}}]),$$

where $\mathbf{H}_{Y_u}$ and $\mathbf{H}_{\mathcal{C}_{Y_u}}$ denote the representations of the target utterance Y_u and $\mathcal{C}_{Y_u}$, respectively. Concretely, $\mathbf{H}_{Y_u}$ 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 $\mathcal{C}_{Y_u}$. $\mathbf{W}_n$ is the trainable parameter of the NUD classifier and the bias term is omitted for simplicity.

2.4 Training Objective

With the main chat translation task and three auxiliary tasks, the total training objective of the conventional multi-task learning is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{NCT}} + \alpha(\mathcal{L}_{\text{MRG}} + \mathcal{L}_{\text{XRG}} + \mathcal{L}_{\text{NUD}}), \quad (2)$$

where α is the balancing factor between $\mathcal{L}_{\text{NCT}}$ and other auxiliary objectives.

3 Scheduled Multi-task Learning for NCT

In this section, we introduce the proposed **S**cheduled **M**ulti-task **L**earning (SML) framework, including three stages: general pre-training, in-domain pre-training, and in-domain fine-tuning, as shown in Fig. 1. Specifically, we firstly describe the process of *in-domain pre-training* (§ 3.1) and then present some *findings of conventional multi-task learning* (§ 3.2), which inspire us to investigate the *scheduled multi-task learning* (§ 3.3). Finally, we

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

elaborate on the process of *training and inference* (§ 3.4).

3.1 In-domain Pre-training

For the second in-domain pre-training, we firstly build an in-domain paired bilingual dialogue data and then conduct pre-training on it.

To construct the paired bilingual dialogue data, we firstly crawl the in-domain consecutive movie subtitles of En↔Zh and download the consecutive movie subtitles of En↔De on related websites[6]. Since both bilingual movie subtitles are not strictly aligned, we utilize the Vecalign tool (Thompson and Koehn, 2019), an accurate sentence alignment algorithm, to align them. Meanwhile, we leverage the LASER toolkit[7] to obtain the multilingual embedding for better alignment performance. Consequently, we obtain two relatively clean paired movie subtitles. According to the setting of dialogue context length in Liang et al. (2021a), we take four consecutive utterances as one dialogue, and then filter out duplicate dialogues. Finally, we attain two in-domain paired bilingual dialogue dataset, the statistics of which are shown in Tab. 1.

Datasets	#Dialogues	#Utterances	#Sentences
En↔Zh	28,214,769	28,238,877	22,244,006
En↔De	18,041,125	18,048,573	45,541,367

Table 1: Statistics of our constructed chat translation data. The **#Sentences** column is the general-domain WMT sentence pairs used in the first pre-training stage.

Based on the constructed in-domain bilingual corpus, we continue to pre-train the NCT model after the general pre-training stage, and then go to the in-domain fine-tuning stage, as shown in the In-domain Pre-training&Fine-tuning parts of Fig. 1.

3.2 Findings of Conventional Multi-task Learning

According to the finding that multi-task learning can enhance the NCT model (Liang et al., 2021d), in the last two training processes (*i.e.*, the In-domain Pre-training and In-domain Fine-tuning parts of Fig. 1), we conduct extensive multi-task learning experiments, aiming to achieve a better NCT model. Firstly, we present one additional auxiliary task, *i.e.* Cross-lingual NUD (XNUD), given the intuition that more dialogue-related tasks may yield better performance. Then, we conclude some multi-task learning findings that could motivate us to investigate how to use these auxiliary tasks well.

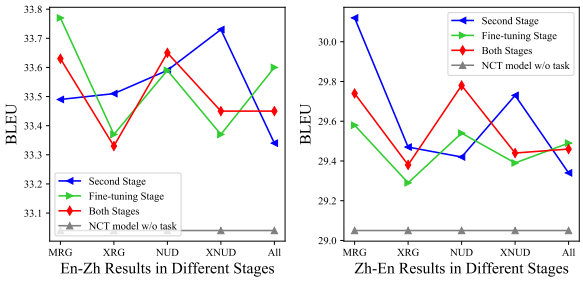


Figure 2: The effect of each task on validation sets in different training stages, under transformer *Base* setting, where "All" denotes all four auxiliary tasks. We find that each auxiliary task performs well on the second stage while XRG and XNUD tasks perform relatively poorly in the fine-tuning stage. Further, we observe that all auxiliary tasks in a conventional multi-task learning manner do not obtain significant cumulative benefits. That is, the auxiliary tasks are *different strokes for different folks*.

XNUD. Similar to the NUD task described in § 2.3, the XNUD aims to distinguish whether the translated text is coherent to be the next utterance of the given cross-lingual dialogue history context. Compared to the NUD task, the different point lies in the cross-lingual dialogue context history, *i.e.*, a positive sample $(\mathcal{C}_{X_u}, Y_{u^+})$ with the label $\ell = 1$ and a negative sample $(\mathcal{C}_{X_u}, Y_{u^-})$ with the label $\ell = 0$. Formally, the training objective of XNUD is defined as follows:

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

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

where $\mathbf{H}_{\mathcal{C}_{X_u}}$ denotes the representation of $\mathcal{C}_{Y_u}$, which is calculated as same as $\mathbf{H}_{\mathcal{C}_{Y_u}}$ in NUD. $\mathbf{W}_x$ is the trainable parameter of the XNUD classifier and the bias term is omitted for simplicity.

Findings. Based on four auxiliary tasks (MRG, XRG, NUD, and XNUD), we investigate in which stage in Fig. 1 the auxiliary tasks work well in a conventional multi-task learning manner[8] and the following is what we find from Fig. 2:

- Each auxiliary task can always bring improvement compared with the NCT model *w/o* task;

[6]En↔Zh: https://www.kexiaoguo.com/ and En↔De: https://opus.nlpl.eu/OpenSubtitles.php

[7]https://github.com/facebookresearch/LASER

[8]Note that, in the last two in-domain stages, we use the conventional multi-task learning to pre-train and fine-tune models rather than the scheduled multi-task learning.

- By contrast, XRG and XNUD tasks perform relatively poorly in the final fine-tuning stage than MRG and NUD tasks;
- Some tasks used only in one stage (*e.g.*, XRG and XNUD in the second stage) perform better than being used in both stages, revealing that different auxiliary tasks may prefer different stages to exert their advantages; (one best setting seems that all tasks are used in the second stage while only MRG and NUD tasks are used in the final fine-tuning stage.)
- Using all auxiliary tasks in a conventional multi-task learning manner does not obtain significant cumulative benefits.

Given the above findings, we wonder whether there exists a strategy to dynamically schedule them to exert their potential for the main NCT task.

3.3 Scheduled Multi-task Learning

Inspired by Yu et al. (2020), we design a gradient-based scheduled multi-task learning algorithm to dynamically schedule all auxiliary tasks at each training step, as shown in Algorithm 1. Specifically, at each training step (**line 1**), for each task we firstly compute its gradient to model parameters θ (**lines 2∼4**, and we denote the gradient of the main NCT task as $\mathbf{g}_{nct}$). Then, we obtain the projection of the gradient $\mathbf{g}_k$ of each auxiliary task k onto $\mathbf{g}_{nct}$ (**line 5**), as shown in Fig. 3. Finally, we utilize the sum of $\mathbf{g}_{nct}$ and all projection (*i.e.*, the blue arrows part, as shown in Fig. 3) of auxiliary tasks to update model parameters.

The core ideas behind the gradient-based SML algorithm are: (1) when the cosine similarity between $\mathbf{g}_k$ and $\mathbf{g}_{nct}$ is positive, *i.e.*, the gradient projection $\mathbf{g}'_k$ is in the same gradient descent direction with the main NCT task, *i.e.*, Fig. 3 (a), which could help the NCT model achieve optimal solution; (2) when the cosine similarity between $\mathbf{g}_k$ and $\mathbf{g}_{nct}$ is negative, *i.e.*, Fig. 3 (b), which can avoid the model being optimized too fast and overfitted. Therefore, we also keep the inverse gradient to prevent the NCT model from overfitting as a regularizer. In this way, such auxiliary task joins in training at each step with the NCT task when its gradient projection is in line with $\mathbf{g}_{nct}$, which acted as a fine-grained joint training manner.

3.4 Training and Inference

Our training process includes three stages: the first pre-training stage on the general-domain sentence pairs (X, Y):

$$\mathcal{L}_{\text{Sent-NMT}} = -\sum_{t=1}^{|Y|} \log(p(y_t|X, y_{<t})), \quad (3)$$

the second in-domain pre-training stage, and the final in-domain fine-tuning stage on the chat translation data:

$$\mathcal{J} = \mathcal{L}_{\text{NCT}} + \alpha \sum_{k}^{\mathcal{T}} \mathcal{L}_k, \quad (4)$$

where $\mathcal{T}$ is the auxiliary tasks set and we keep the balancing hyper-parameter α. Although the form of $\mathcal{L}_k$ is the same with Eq. 2, the gradient that participates in updating model parameters is different where it depends on the gradient descent direction of the NCT task in Eq. 4.

At inference, all auxiliary tasks are not involved and only the NCT model after scheduled multi-task fine-tuning is applied to chat translation.

Algorithm 1: Gradient-based SML

Require: Model parameters θ, Balancing factor α, MaxTrainStep T, NCT task, Auxiliary tasks set $\mathcal{T} = \{\text{MRG}, \text{XRG}, \text{NUD}, \text{XNUD}\}$.

Init: θ, $t = 0$

1 **for** $t < T$ **do**

2 $\quad \mathbf{g}_{nct} \leftarrow \nabla_\theta \mathcal{L}_{\text{NCT}}(\theta)$

3 $\quad$ **for** k *in* $\mathcal{T}$ **do**

4 $\quad\quad \mathbf{g}_k \leftarrow \nabla_\theta \mathcal{L}_k(\theta)$

5 $\quad\quad$ Set $\mathbf{g}'_k = \frac{\mathbf{g}_k \cdot \mathbf{g}_{nct}}{\|\mathbf{g}_{nct}\|^2} \mathbf{g}_{nct}$

Return: Update $\Delta\theta = \mathbf{g}_{nct} + \alpha \sum_k \mathbf{g}'_k$

(a) (b)

Figure 3: Gradient projection example.

4 Experiments

4.1 Datasets and Metrics

Datasets. The training of our SML framework consists of three stages: (1) pre-train the model on a large-scale sentence-level NMT corpus (WMT20[9]);

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

	Models	En→Zh		Zh→En		En→De		De→En	
		BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓
Base	Trans. *w/o* FT	21.40	72.4	18.52	59.1	40.02	42.5	48.38	33.4
	Trans.	25.22	62.8	21.59	56.7	58.43	26.7	59.57	26.2
	Dia-Trans.	24.96	63.7	20.49	60.1	58.33	26.8	59.09	26.2
	Gate-Trans.	25.34	62.5	21.03	56.9	58.48	26.6	59.53	26.1
	NCT	24.76	63.4	20.61	59.8	58.15	27.1	59.46	25.7
	CPCC	27.55	60.1	22.50	55.7	60.13	25.4	61.05	24.9
	CSA-NCT	27.77	60.0	22.36	55.9	59.50	25.7	60.65	25.4
	SML (Ours)	**32.25**††	**55.1**††	**26.42**††	**51.4**††	**60.65**†	**25.3**	**61.78**††	**24.6**†
Big	Trans. *w/o* FT	22.81	69.6	19.58	57.7	40.53	42.2	49.90	33.3
	Trans.	26.95	60.7	22.15	56.1	59.01	26.0	59.98	25.9
	Dia-Trans.	26.72	62.4	21.09	58.1	58.68	26.8	59.63	26.0
	Gate-Trans.	27.13	60.3	22.26	55.8	58.94	26.2	60.08	25.5
	NCT	26.45	62.6	21.38	57.7	58.61	26.5	59.98	25.4
	CPCC	28.98	59.0	22.98	54.6	60.23	25.6	61.45	24.8
	CSA-NCT	28.86	58.7	23.69	54.7	60.64	25.3	61.21	24.9
	SML (Ours)	**32.87**††	**54.4**††	**27.58**††	**50.6**††	**61.16**†	**25.0**†	**62.17**††	**24.4**†

Table 2: Test results on BMELD (En↔Zh) and BConTrasT (En↔De) in terms of BLEU (%) and TER (%). The best and second best 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 $p < 0.01$ hereinafter, respectively. The results of contrast models are from Liang et al. (2021a,d). Strictly speaking, it is unfair to directly compare with them since we use additional data. Therefore, we conduct further experiments in Tab. 3 for fair comparison.

(2) further pre-train the model on our constructed in-domain chat translation corpus; (3) fine-tune on the target chat translation corpus: BMELD (Liang et al., 2021a) and BConTrasT (Farajian et al., 2020). The target dataset details (*e.g.*, splits of training, validation or test sets) are described in Appendix A.

Metrics. Following Liang et al. (2021d), we use SacreBLEU[10] (Post, 2018) and TER (Snover et al., 2006) with the statistical significance test (Koehn, 2004) for fair comparison. Specifically, we report character-level BLEU for En→Zh, case-insensitive BLEU score for Zh→En, and case-sensitive BLEU score likewise for En↔De.

4.2 Implementation Details

In this paper, we adopt the settings of standard *Transformer-Base* and *Transformer-Big* in Vaswani et al. (2017). Generally, we utilize the settings in Liang et al. (2021d) for fair comparison. For more details, please refer to Appendix B. We investigate the effect of the XNUD task in § 5.4, where the new XNUD performs well based on existing auxiliary tasks.

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

4.3 Comparison Models

Sentence-level NMT Systems. Trans. *w/o* FT and Trans. (Vaswani et al., 2017): both are the de-facto transformer-based NMT models, and the difference is that the "Trans." model is fine-tuned on the chat translation data after being pre-trained on sentence-level NMT corpus.

Context-aware NMT Systems. Dia-Trans. (Maruf et al., 2018): A Transformer-based model where an additional encoder is used to introduce the mixed-language dialogue history, re-implement by Liang et al. (2021a).

Gate-Trans. (Zhang et al., 2018) and NCT (Ma et al., 2020): Both are document-level NMT Transformer models where they introduce the dialogue history by a gate and by sharing the first encoder layer, respectively.

CPCC (Liang et al., 2021a): A variational model that focuses on incorporating dialogue characteristics into a translator for better performance.

CSA-NCT (Liang et al., 2021d): A multi-task learning model that uses several auxiliary tasks to help generate dialogue-related translations.

	Models (*Base*)	En→Zh		Zh→En	
		BLEU↑	TER↓	BLEU↑	TER↓
Two-stage *w/o* data	Trans. *w/o* FT	21.40	72.4	18.52	59.1
	Trans.	25.22	62.8	21.59	56.7
	NCT	24.76	63.4	20.61	59.8
	M-NCT	27.84	59.8	22.41	55.9
	SML (Ours)	**28.96**††	**58.3**††	**23.23**††	**55.2**††
Three-stage *w/* data	Trans. *w/o* FT	28.60	56.7	22.46	53.9
	Trans.	30.90	56.5	25.04	53.3
	NCT	31.37	55.9	25.35	52.7
	M-NCT	31.63	55.6	25.86	51.9
	SML (Ours)	**32.25**††	**55.1**††	**26.42**†	**51.4**††

Table 3: Results on test sets of BMELD in terms of BLEU (%) and TER (%), where "Two-stage *w/o* data" means the pre-training-then-fine-tuning paradigm and the in-domain data not being used, and "Three-stage *w/* data" means the proposed three-stage method and this group uses the in-domain data. The "M-NCT" denotes the multi-task learning model jointly trained with four auxiliary tasks in a conventional manner. All models apply the same two/three-stage training strategy with our SML model for fair comparison except the "Trans. *w/o* FT" model, respectively.

4.4 Main Results

In Tab. 2, We report the main results on En↔Zh and En↔De under *Base* and *Big* settings. In Tab. 3, we present additional results on En↔Zh.

Results on En↔Zh. Under the *Base* setting, our model significantly outperforms the sentence-level/context-aware baselines by a large margin (*e.g.*, the previous best "CSA-NCT"), 4.58↑ on En→Zh and 4.06↑ on Zh→En, showing the effectiveness of the large-scale in-domain data and our scheduled multi-task learning. In terms of TER, SML also performs best on the two directions, 5.0↓ and 4.3↓ than "CPCC" (the lower the better), respectively. Under the *Big* setting, our model consistently surpasses all existing systems once again.

Results on En↔De. On both En→De and De→En, our model presents notable improvements over all comparison models by up to 2.50↑ and 2.69↑ BLEU gains under the *Base* setting, and by 2.55↑ and 2.53↑ BLEU gains under the *Big* setting, respectively. These results demonstrate the superiority of our three-stage training framework and also show the generalizability of our model across different language pairs. Since the baselines of En↔De are very strong, the results of En↔De are not so significant than En↔Zh.

#	Where to Use?	En→Zh		Zh→En	
		BLEU↑	TER↓	BLEU↑	TER↓
0	Two-stage (Not Use)	29.49	55.8	24.15	53.3
1	Two-stage (①)	31.17	53.2	26.14	51.4
2	Two-stage (②)	29.87	53.7	27.47	50.5
3	Three-stage (②)	**33.45**††	**51.1**††	**29.47**††	**49.3**††

Table 4: Results on validation sets of where to use the large-scale in-domain data under the *Base* setting. The rows 0∼2 use the pre-training-then-fine-tuning (*i.e.*, two-stage) paradigm while row 3 is the proposed three-stage method. For a fair comparison, the final fine-tuning stage of rows 0∼3 is all trained in the conventional multi-task training manner and the only difference is the usage of the in-domain data. Specifically, row 0 denotes without using the in-domain data. Row 1 denotes that we incorporate the in-domain data into the first pre-training stage (①). Row 2 denotes that we introduce the in-domain data into the fine-tuning stage (②). Row 3 denotes that we add a second pre-training stage to introduce the in-domain data.

Additional Results. Tab. 2 presents our overall model performance, though, strictly speaking, it is unfair to directly compare our approaches with previous ones. Therefore, we conduct additional experiments in Tab. 3 under two settings: (*i*) using the original pre-training-then-fine-tuning framework without introducing the large-scale in-domain data (*i.e.*, "Two-stage *w/o* data" group); (*ii*) using the proposed three-stage method with the large-scale in-domain data (*i.e.*, "Three-stage *w/* data" group). And we conclude that (1) the same model (*e.g.*, SML) can be significantly enhanced by the second in-domain pre-training stage, demonstrating the effectiveness of the second pre-training on the in-domain data; (2) our SML model always exceeds the conventional multi-task learning model "M-NCT" in both settings, indicating the superiority of the scheduled multi-task learning strategy.

5 Analysis

5.1 Ablation Study

We conduct ablation studies in Tab. 4 and Tab. 5 to answer the following two questions. **Q1**: ***why a three-stage training framework?*** and **Q2**: ***why the scheduled multi-task learning strategy?***

To answer **Q1**, in Tab. 4, we firstly investigate the effect of the large-scale in-domain chat translation data and further explore where to use it. Firstly, the results of rows 1∼3 substantially outperform those in row 0, proving the availability of incorporating the in-domain data. Secondly, the results of

#	Training Manners?	En→Zh		Zh→En	
		BLEU↑	TER↓	BLEU↑	TER↓
0	Conventional Multi-task Learning	33.45	51.2	29.47	49.3
1	Random Multi-task Learning	32.88	51.6	29.19	49.5
2	Prior-based Multi-task Learning	33.94	51.1	29.74	49.1
3	Scheduled Multi-task Learning (SML)	**34.21**†	**51.0**	**30.13**†	**49.0**
4	SML *w/o* inverse gradient projection	33.85	51.1	29.79	49.1

Table 5: Results on validation sets of the three-stage training framework in different multi-task training manners, under the *Base* setting. Row 1 denotes that the auxiliary tasks are randomly added in a conventional training manner at each training step. Row 2 denotes that we add the auxiliary tasks according to their performance in different stages, *i.e.*, we add all tasks in the second stage while only considering MRG and NUD in the fine-tuning stage according to prior trial results in Fig. 2. Row 4 denotes that we remove the inverse gradient projection of auxiliary tasks (*i.e.*, Fig. 3 (b)).

row 3 significantly surpass rows 1∼2, indicating that the in-domain data used in the proposed second stage of our three-stage training framework is very successful rather than used in the stage of pre-training-then-fine-tuning paradigm. That is, the experiments show the effectiveness and necessity of our three-stage training framework.

To answer **Q2**, we investigate multiple multi-task learning strategies in Tab. 5. Firstly, the results of row 3 are notably higher than those of rows 0∼2 in both language directions, obtaining significant cumulative benefits of auxiliary tasks than rows 0∼2, demonstrating the validity of the proposed SML strategy. Secondly, the results of row 3 vs row 4 show that the inverse gradient projection of auxiliary tasks also has a positive impact on the model performance, which may prevent the model from overfitting, working as a regularizer. All experiments show the superiority of our scheduled multi-task learning strategy.

5.2 Human Evaluation

Inspired by Bao et al. (2020) and Liang et al. (2021a), we use two criteria for human evaluation to judge whether the translation is:

1. semantically **coherent** with the dialogue history?
2. **fluent** and grammatically correct?

Firstly, we randomly sample 200 conversations from the test set of BMELD in En→Zh. Then, we use 6 models in Tab. 6 to generate translated utterances of these sampled conversations. Finally, we assign the translated utterances and their corresponding dialogue history utterances in the target language to three postgraduate human annotators, and then ask them to make evaluations (0/1 score) according to the above two criteria, and average the scores as the final result.

Models (*Base*)	**Coherence**	**Fluency**
Trans. *w/o* FT	0.585	0.630
Trans.	0.620	0.655
NCT	0.635	0.665
CSA-NCT	0.650	0.680
M-NCT	0.665	0.695
SML (Ours)	**0.690**†	**0.735**†

Table 6: Results of human evaluation (En→Zh). All models use the three-stage training framework to introduce the in-domain data.

Models (*Base*)	**1-th Pr.**	**2-th Pr.**	**3-th Pr.**
Trans. *w/o* FT	58.11	55.15	52.15
Trans.	58.77	56.10	52.71
NCT	59.19	56.43	52.89
CSA-NCT	59.45	56.74	53.02
M-NCT	59.57	56.79	53.18
SML (Ours)	60.48††	57.88††	53.95††
Human Reference	**61.03**	**59.24**	**54.19**

Table 7: Results (%) of dialogue coherence in terms of sentence similarity on validation set of BMELD in En→Zh direction. 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 ($p < 0.01$). All models use the three-stage training framework to introduce the in-domain data.

Tab. 6 shows that our model generates more coherent and fluent translations when compared with other models (significance test, $p < 0.05$), which shows the superiority of our model. The inter-annotator agreements calculated by the Fleiss' kappa (Fleiss and Cohen, 1973) are 0.558 and 0.583 for coherence and fluency, respectively. It indicates "Moderate Agreement" for both criteria.

5.3 Dialogue Coherence

We measure dialogue coherence as sentence similarity following Lapata and Barzilay (2005); Xiong et al. (2019); Liang et al. (2021a):

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

where cos denotes cosine similarity and $f(s_i) = \frac{1}{|s_i|}\sum_{\mathbf{w}\in s_i}(\mathbf{w})$ and $\mathbf{w}$ is the vector for word w, and

Models (*Base*)	En→Zh		Zh→En	
	BLEU↑	TER↓	BLEU↑	TER↓
NCT+{MRG,CRG,NUD}	28.94	56.0	23.82	54.3
NCT+{MRG,CRG,NUD,XNUD}	**29.49**††	**55.8**	**24.15**†	**53.5**††

Table 8: The results on validation sets after adding the XNUD task on three auxiliary tasks, *i.e.*, MRG, XRG and NUD (Liang et al., 2021d), which are trained in conventional manner (without incorporating in-domain data).

s_i is the sentence. We use Word2Vec[11] (Mikolov et al., 2013) trained on a dialogue dataset[12] to obtain the distributed word vectors whose dimension is set to 100.

Tab. 7 shows the measured coherence of different models on validation set of BMELD in En→Zh direction. It shows that our SML produces more coherent translations compared to all existing models (significance test, $p < 0.01$).

5.4 Effect of the Auxiliary Task: XNUD

We investigate the effect of the XNUD task. As shown in Tab. 8, the "M-NCT" denotes the multi-task learning model jointly trained with four auxiliary tasks in conventional manner. After removing the XNUD task, the performance drops to some extend, indicating that the new XNUD task achieves further performance improvement based on three existing auxiliary tasks (Liang et al., 2021d). Then, based on the strong "M-NCT" model, we further investigate where and how to make the most of them for the main NCT task.

6 Related Work

Neural Chat Translation. The goal of NCT is to train a dialogue-aware translation model using the bilingual dialogue history, which is different from document-level/sentence-level machine translation (Maruf et al., 2019; Ma et al., 2020; Yan et al., 2020; Meng and Zhang, 2019; Zhang et al., 2019). Previous work can be roughly divided into two categories. One (Wang et al., 2016b; Maruf et al., 2018; Zhang and Zhou, 2019; Rikters et al., 2020) mainly pays attention to automatically constructing the bilingual corpus since no publicly available human-annotated data (Farajian et al., 2020). The other (Wang et al., 2021; Liang et al., 2021a,d) aims to incorporate the bilingual dialogue characteristics into the NCT model via multi-task learning. Different from the above studies, we focus on introducing the in-domain chat translation data to learn domain-specific patterns and scheduling the auxiliary tasks to exert their potential for high translation quality.

Multi-task Learning. Conventional multi-task learning (MTL) (Caruana, 1997), which trains the model on multiple related tasks to promote the representation learning and generalization performance, has been successfully used in many NLP tasks (Collobert and Weston, 2008; Ruder, 2017; Deng et al., 2013; Liang et al., 2021c,b). In the NCT, conventional MTL has been explored to inject the dialogue characteristics into models with dialogue-related tasks such as response generation (Liang et al., 2021a,d). In this work, we instead focus on how to schedule the auxiliary tasks at training to make the most of them for better translations.

7 Conclusion

This paper proposes a scheduled multi-task learning framework armed with an additional in-domain pre-training stage and a gradient-based scheduled multi-task learning strategy. Experiments on En↔Zh and En↔De demonstrate that our framework significantly improves translation quality on both BLEU and TER metrics, showing its effectiveness and generalizability. Human evaluation further verifies that our model yields better translations in terms of coherence and fluency. Furthermore, we contribute two large-scale in-domain paired bilingual dialogue datasets to the research community.

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). Liang is supported by 2021 Tencent Rhino-Bird Research Elite Training Program. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

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

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

[12] We choose our constructed dialogue corpus to learn the word embedding.

neural machine translation. In *Proceedings of WMT*, pages 454–459.

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.

Rich Caruana. 1997. Multitask learning. In *Machine Learning*, pages 41–75.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, page 160–167.

Li Deng, Geoffrey E. Hinton, and Brian Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. *2013 IEEE ICASSP*, pages 8599–8603.

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.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. Train no evil: Selective masking for task-guided pre-training. In *Proceedings of EMNLP*, pages 6966–6974.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, pages 8342–8360.

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.

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

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.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. Msctd: A multimodal sentiment chat translation dataset. *arXiv preprint arXiv:2202.13645*.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021b. A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. *Neurocomputing*.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021c. An iterative multi-knowledge transfer network for aspect-based sentiment analysis. In *Findings of EMNLP*, pages 1768–1780.

Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2021d. Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of EMNLP*, pages 67–79.

Tongtong Liu, Fangxiang Feng, and Xiaojie Wang. 2021. Multi-stage pre-training over simplified multimodal pre-training models. In *Proceedings of ACL*, pages 2556–2565.

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*.

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.

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.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of AACL*, pages 557–575.

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.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of ACL*, pages 5231–5247.

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.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

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*.

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.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of EMNLP*, pages 1342–1348.

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

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

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016b. 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. In *Proceedings of NAACL: Human Language Technologies: Industry Papers*, pages 105–112.

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.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of EMNLP-IJCNLP*, pages 1527–1537.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Proceedings of NIPS*, volume 33, pages 5824–5836.

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.

A Datasets

As mentioned in § 4.1, our experiments involve the WMT20 dataset for general-domain pre-training, the newly constructed in-domain chat translation data for the second pre-training (please refer to § 3.1), and two target chat translation corpora, BMELD (Liang et al., 2021a) and BConTrasT (Farajian et al., 2020). The statistics about the splits of training, validation, and test sets of BMELD (En↔Zh) and BConTrasT (En↔De) are shown in Tab. 9.

WMT20. Following Liang et al. (2021a,d), For En↔Zh, we combine News Commentary v15, Wiki Titles v2, UN Parallel Corpus V1.0, CCMT Corpus, and WikiMatrix. For En↔De, we combine six corpora including Euporal, ParaCrawl, CommonCrawl, TildeRapid, NewsCommentary, 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 22,244,006 sentence pairs for En↔Zh and 45,541,367 sentence pairs for En↔De, respectively.

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)[13], 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 CoreNLP toolkit[14].

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

Table 9: Statistics of chat translation data.

BConTrasT. The dataset[15] 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[16] 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.

B Implementation Details

For all experiments, we follow the settings of Vaswani et al. (2017), namely *Transformer-Base* and *Transformer-Big*. 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. For fair comparison, we set the training step for the first pre-training stage and the second pre-training stage totally to 200,000 (100,000 for each stage), and

[13]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.

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

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

[16]www.unbabel.com

set the step of fine-tuning stage 5,000. As for the balancing factor α in Eq. 4, we follow (Liang et al., 2021d) to decay α from 1 to 0 over training steps (we set them to 100,000 and 5,000 for the last two training stages, respectively). The batch size for each GPU is set to 4096 tokens. All experiments in three stages are conducted utilizing 8 NVIDIA Tesla V100 GPUs, which gives us about 8*4096 tokens per update for all experiments. All models are optimized using Adam (Kingma and Ba, 2014) 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. $|T|$ is set to 10. When building the shared vocabulary $|V|$, we keep such word if its frequency is larger than 100. 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.

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), which is similar for other translation directions.

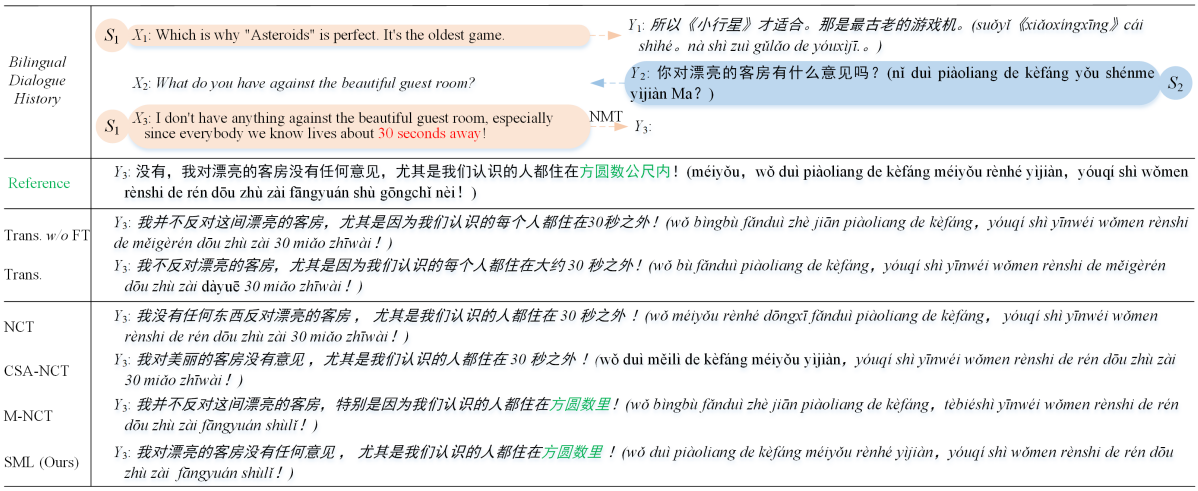
C Case Study

In this section, we present two illustrative cases in Fig. 4 to give some observations among the comparison models and ours.

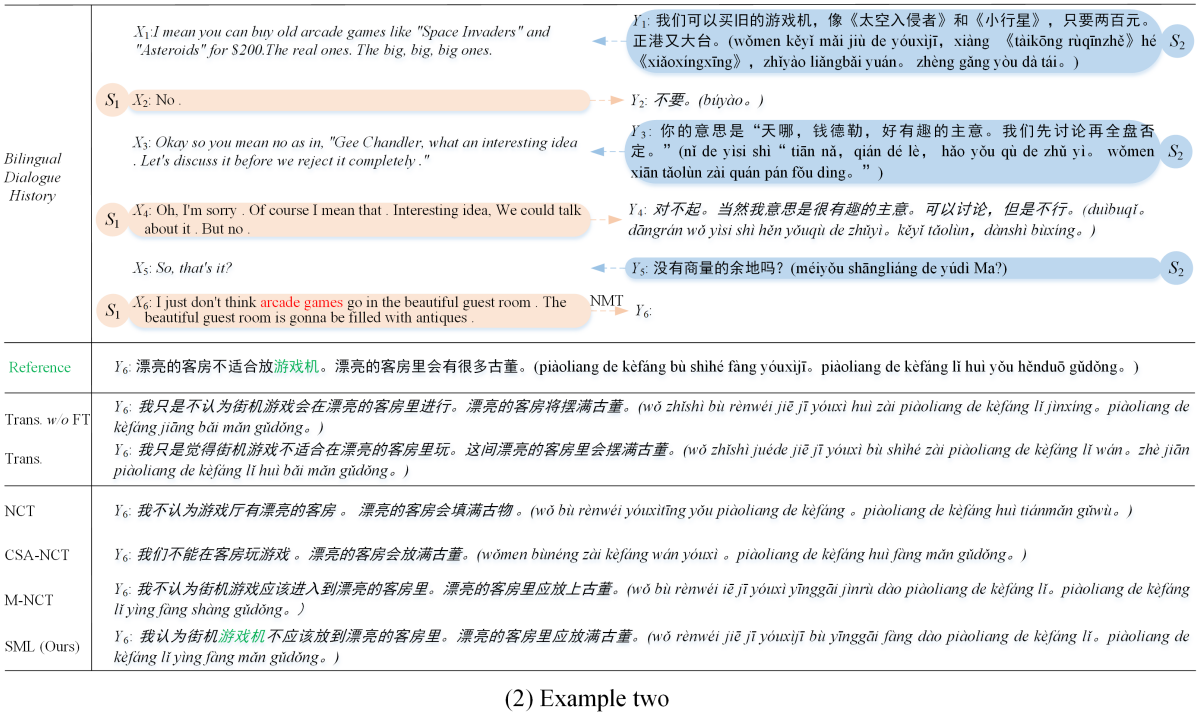
For the case Fig. 4 (1), we find that most comparison models just translate the phrase "30 seconds away" literally as "30 秒之外 (30 *miǎo zhīwài*)", which is very strange and is not in line with Chinese language habits. By contrast, the "M-NCT" and "SML" models, through three-stage training, capture such translation pattern and generate an appropriate Chinese phrase "方圆数里 (fāngyúan shùlǐ)". The reason behind this is that the large-scale in-domain dialogue bilingual corpus contains many cases of free translation, which is common in daily conversations translation. This suggests that the in-domain pre-training is indispensable for a successful chat translator.

For the case Fig. 4 (2), we find that all comparison models fail to translate the word "games", where they translate it as "游戏 (yóuxì)". The reason may be that they cannot fully understand the dialogue context even though some models (*e.g.*, "CSA-NCT" and "M-NCT") also jointly trained with the dialogue-related auxiliary tasks. By contrast, the "SML" model, enhanced by multi-stage scheduled multi-task learning, obtains accurate results.

In summary, the two cases show that our SML model enhanced by the in-domain data and scheduled multi-task learning yields satisfactory translations, demonstrating its effectiveness and superiority.



(1) Example one



(2) Example two

Figure 4: The illustrative cases of bilingual conversation translation.