MSCTD: A Multimodal Sentiment Chat Translation Dataset

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

Multimodal machine translation and textual chat translation have received considerable attention in recent years. Although the conversation in its natural form is usually multimodal, there still lacks work on multimodal machine translation in conversations. In this work, we introduce a new task named **M**ultimodal **C**hat **T**ranslation (MCT), aiming to generate more accurate translations with the help of the associated dialogue history and visual context. To this end, we firstly construct a **M**ultimodal **S**entiment **C**hat **T**ranslation **D**ataset (MSCTD) containing 142,871 English-Chinese utterance pairs in 14,762 bilingual dialogues and 30,370 English-German utterance pairs in 3,079 bilingual dialogues. Each utterance pair, corresponding to the visual context that reflects the current conversational scene, is annotated with a sentiment label. Then, we benchmark the task by establishing multiple baseline systems that incorporate multimodal and sentiment features for MCT. Preliminary experiments on four language directions (English↔Chinese and English↔German) verify the potential of contextual and multimodal information fusion and the positive impact of sentiment on the MCT task. Additionally, as a by-product of the MSCTD, it also provides two new benchmarks on multimodal dialogue sentiment analysis. Our work can facilitate research on both multimodal chat translation and multimodal dialogue sentiment analysis.[1]

1 Introduction

Multimodal machine translation (Huang et al., 2016; Calixto and Liu, 2017) and textual chat translation (Wang et al., 2016; Farajian et al., 2020; Liang et al., 2021a) mainly focus on investigating the potential visual features and dialogue context, respectively. Both of them have received much attention. Although plenty of studies on them have been carried out based on either image captions (Calixto et al., 2017, 2019; Ive et al., 2019; Yin et al., 2020; Yao and Wan, 2020) or textual dialogues (Wang et al., 2017; Maruf et al., 2018; Liang et al., 2021c), to our knowledge, little research work has been devoted to multimodal machine translation in conversations. One important reason is the lack of multimodal bilingual conversational datasets.

Generally, conversation in its natural form is multimodal (Poria et al., 2019; Liang et al., 2021b). When humans converse, what a speaker would say next depends largely on what he/she sees. That is, the visual information plays a key role in (*i*) supplementing some crucial scene information (*e.g.*, the specific locations or objects, or facial expressions), (*ii*) resolving ambiguous multi-sense words (*e.g.*, bank), and (*iii*) addressing pronominal anaphora issues (*e.g.*, it/this). For instance, as shown in Fig. 1 (a), the image obviously points out the current location "on the sea", which may help disambiguate the meaning of "course" in the utterance X_5. Specifically, the dialogue history (*i.e.*, talking about maritime affairs) and the corresponding visual context (*i.e.*, on the sea/boat) assist us to determine that the word "course" means "route/direction" instead of "curriculum". In Fig. 1 (b), the visual context indicates object information, *i.e.*, the "defibrillator" in X_1, which may help with translation. In Fig. 1 (c), the image of the utterance X_1 also demonstrates that it can provide appropriate candidates (*i.e.*, the jeans) when translating the pronoun "these". Besides, the image offers some clues to judge the sentiment when it is hard to judge the polarity based only on the utterance (*e.g.*, Y_2 in Fig. 1 (b) and X_3 in Fig. 1 (c)). All of the above call for a real-life multimodal bilingual conversational data resource that can encourage further research in chat transla-
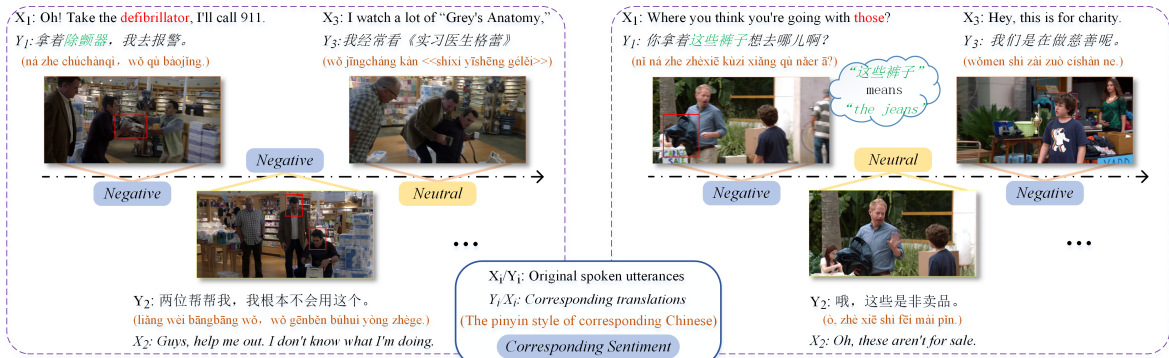

(a) The current location "on the sea" helps disambiguate polysemy ("course" in X_5).

(b) The red framed object in the image of X_1 may help the translation of "defibrillator" in X_1.

(c) The red framed object (means "jeans") in the image of X_1 may help the translation of the pronoun "those" in X_1.

Figure 1: Three examples of the annotated multimodal bilingual dialogue in our MSCTD and the conversation is going from left to right.

tion.

In this work, we propose a new task named **Multimodal Chat Translation** (MCT), with the goal to produce more accurate translations by taking the dialogue history and visual context into consideration. To this end, we firstly construct a **Multimodal Sentiment Chat Translation Dataset** (MSCTD). The MSCTD includes over 17k multimodal bilingual conversations (more than 142k English-Chinese and 30k English-German utterance pairs), where each utterance pair corresponds with the associated visual context indicating where it happens. In addition, each utterance is annotated with one sentiment label (*i.e.*, positive/neutral/negative).

Based on the constructed MSCTD, we benchmark the MCT task by establishing multiple Transformer-based (Vaswani et al., 2017) systems adapted from several advanced representative multimodal machine translation models (Ive et al., 2019; Yao and Wan, 2020) and textual chat translation models (Ma et al., 2020; Liang et al., 2021c). Specifically, we incorporate multimodal features and sentiment features into these mod-

els for a suitable translation under the current conversational scene. Extensive experiments on four language directions (English↔Chinese and English↔German) in terms of BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and TER (Snover et al., 2006), demonstrate the effectiveness of contextual and multimodal information fusion, and the positive impact of sentiment on MCT. Furthermore, experiments on the multimodal dialogue sentiment analysis task of the three languages show the added value of the proposed MSCTD.

In summary, our main contributions are:

- We propose a new task: multimodal chat translation named MCT, to advance multimodal chat translation research.

- We are the first that contributes the human-annotated multimodal sentiment chat translation dataset (MSCTD), which contains 17,841 multimodal bilingual conversations, totally 173,240 <English utterance, Chinese/German utterance, image, sentiment> quadruplets.

- We implement multiple Transformer-based baselines and provide benchmarks for the new

2602

task. We also conduct comprehensive analysis and ablation study to offer more insights.

- As a by-product of our MSCTD, it also facilitates the development of multimodal dialogue sentiment analysis.

2 Tasks

In this section, we firstly clarify the symbol definition, and then define the proposed *Multimodal Chat Translation* task and the existing *Multimodal Dialogue Sentiment Analysis* task.

In a multimodal bilingual conversation (*e.g.*, Fig. 1 (a)), we assume the two speakers have alternatively given utterances in different languages for u turns, resulting in $X_1, X_2, X_3, X_4, ..., X_u$ and $Y_1, Y_2, Y_3, Y_4, ..., Y_u$ on the source and target sides, respectively, along with the corresponding visual context representing where it happens: $Z_1, Z_2, Z_3, Z_4, ..., Z_u$. Among these utterances, $X_1, X_3, X_5, ..., X_u$ are originally spoken by the first speaker 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 by the second speaker and $X_2, X_4, X_6, ..., X_{u-1}$ are the translated utterances in the source language. According to languages and modalities, we define three types of context: (1) 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 (2) that of Y_u on the target side as $\mathcal{C}_{Y_u} = \{Y_1, Y_2, Y_3, ..., Y_{u-1}\}$, and (3) the visual dialogue context $\mathcal{C}_{Z_u} = \{Z_1, Z_2, Z_3, ..., Z_{u-1}, Z_u\}$.[2]

Multimodal Chat Translation. When translating the u-th utterance $X_u = \{x_{u,1}, x_{u,2}, ..., x_{u,N}\}$, the goal of the MCT task is to generate $Y_u = \{y_{u,1}, y_{u,2}, ..., y_{u,T}\}$ with the guidance of bilingual dialogue history contexts $\mathcal{C}_{X_u}$ and $\mathcal{C}_{Y_u}$ and the associated visual context $\mathcal{C}_{Z_u}$. Formally, the probability distribution of the target utterance Y_u is defined as follows:

$$P(Y_u|X_u, \mathcal{C}_u) = \prod_{t=1}^{T} p(y_{u,t}|y_{u,<t}, X_u, \mathcal{C}_u), \quad (1)$$

where $y_{u,<t} = \{y_{u,1}, y_{u,2}, y_{u,3}, ..., y_{u,t-1}\}$ and $\mathcal{C}_u = \{\mathcal{C}_{X_u}, \mathcal{C}_{Y_u}, \mathcal{C}_{Z_u}\}$.

---

[2]For each item of $\{\mathcal{C}_{X_u}, \mathcal{C}_{Y_u}\}$, 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).

Multimodal Dialogue Sentiment Analysis. Taking the u-th utterance X_u for example, the task aims to predict a sentiment label $\ell \in \{$*Positive*, *Neutral*, *Negative*$\}$ for it given the corresponding image Z_u and the dialogue history $\mathcal{C}_{X_u}$.

3 Dataset

In this section, we mainly introduce our MSCTD in five aspects: *Data Source* § 3.1, *Annotation Procedure* § 3.2, *Annotation Quality Assessment* § 3.3, *Dataset Statistics* § 3.4, and the introduction of *Related Datasets* § 3.5.

3.1 Data Source

We mainly select the multimodal dialogues from the public available OpenViDial dataset (Meng et al., 2021), where each monolingual (English) utterance corresponds to an image. Since the original English utterance in OpenViDial is automatically extracted from the corresponding movie image by optical character recognition (OCR)[3], it contains a lot of noises or errors. Furthermore, the lack of associated translations and sentiment labels for utterances, makes it impossible for directly conducting research on multimodal chat translation, sentiment-aware machine translation, and multimodal dialogue sentiment analysis with this data. Therefore, we further correct the wrong English utterances and annotate the corresponding Chinese/German translations and sentiment labels.

3.2 Annotation Procedure

To build the MSCTD, the annotation procedure includes two steps: automatic annotation and then human annotation according to the annotation rules.
Automatic Annotation. To improve the annotation efficiency, we firstly construct a paired English-Chinese subtitle database[4]. Then, we utilize the original English utterance to automatically select its Chinese translation by perfectly matching the English subtitle in the constructed bilingual database. As a result, about 78.57% original English utterances are paired with Chinese translations.

---

[3]https://github.com/JaidedAI/EasyOCR

[4]To build this database, we firstly crawl two consecutive English and Chinese movie subtitles (not aligned) from here https://www.kexiaoguo.com/. Then, we use several advanced technologies (*e.g.*, Vecalign (Thompson and Koehn, 2019) and LASER (Schwenk, 2018)) to align these subtitles. Finally, we obtain the large-scale bilingual dialogue dataset (28M). We will also release this dataset, together with the MSCTD, to facilitate subsequent research.

MSCTD	Type	#Dial.	#Utter.	#Images	#AvgTurns	#AvgEn	#AvgZh/De	#Pos.	#Neu.	#Neg.
	Train	13,749	62,593	62,593	9.65	8.35	10.84	16,902	24,074	21,617
Chinese→English	Valid	504	2,389	2,389	10.05	8.27	10.84	708	809	872
	Test	509	2,385	2,385	9.95	8.13	11.09	756	618	1,011
	Train	13,749	70,148	70,148	9.65	8.34	10.84	18,478	27,762	23,908
English→Chinese	Valid	504	2,674	2,674	10.05	8.14	10.93	746	955	973
	Test	509	2,682	2,682	9.95	8.19	10.97	850	680	1,152
	Train	2,066	9,561	9,561	9.80	8.39	8.46	2,581	3,281	3,699
German→English	Valid	504	2,389	2,389	10.05	8.27	8.17	708	809	872
	Test	509	2,385	2,385	9.95	8.13	8.36	756	618	1,011
	Train	2,066	10,679	10,679	9.80	8.40	8.45	2,902	3,640	4,137
English→German	Valid	504	2,674	2,674	10.05	8.14	8.14	746	955	973
	Test	509	2,682	2,682	9.95	8.19	8.29	850	680	1,152
	T./A.	17,841	173,241	173,241	9.91	8.25	10.91/8.31	46,983	64,881	61,376

Table 1: Detailed Statistics of our MSCTD. #: number of the corresponding item, *i.e.*, Dial.: dialogues; Utter.: utterances; AvgTurns: Average turn length of each dialogue; AvgEn: Average length of each turn in English (word level); AvgZh/De: Average length of each turn in Chinese (character level) and in German (word level); Pos./Neu./Neg.: positive/neutral/negative sentiment label. The "T./A." means the Total number or Average value of each column.

Human Annotation. Since the full data are large, we divide the data into three parts and employ three annotators who are Chinese postgraduate students highly proficient in English comprehension. Each annotator is responsible for annotating one part according to the following guidelines:

- Check and correct each English utterance;
- Check and correct the matched Chinese subtitle to suit the current conversational scene;
- For the remaining 21.43% (without Chinese subtitles), translate them according to the corrected English utterance, the corresponding image, and the dialogue history.

Additionally, we employ another three annotators to label sentiment polarity for each utterance independently (*i.e.*, each one annotates the full data) according to the current utterance, the associated image and the dialogue history. Following Firdaus et al. (2020), majority voting scheme is used for selecting the final sentiment label for each utterance.

Finally, having the conversations in both languages allows us to simulate bilingual conversations where one speaker speaks in English and the other responds in Chinese (Farajian et al., 2020; Liang et al., 2021a). Fig. 1 shows three bilingual conversations where the two speakers have alternatively given utterances, along with their corresponding translations. By doing so, we build the MSCTD[5].

3.3 Annotation Quality Assessment

To evaluate the quality of annotation, we use Fleiss' Kappa to measure the overall annotation consistency among three annotators (Fleiss and Cohen, 1973). We measure this data from two aspects: translation quality and sentiment quality.

For translation quality, we measure the inter-annotator agreement on a subset of data (sample 50 dialogues with 504 utterances), and we ask the three annotators mentioned above to re-annotate this subset independently. Then, we invite another postgraduate student to measure the inter-annotator agreement on the re-annotated subset by the three annotators. Finally, the inter-annotator agreement calculated by Fleiss' kappa are 0.921 for English↔Chinese and 0.957 for English↔German, respectively. They indicate "Almost Perfect Agreement" between three annotators.

For sentiment quality, we measure the inter-annotator agreement on the full dataset. The inter-annotator agreements calculated by Fleiss' kappa is 0.695, which indicates "Substantial Agreement" between three annotators. The level is consistent with previous work (Firdaus et al., 2020) which can be considered as reliable.

3.4 Dataset Statistics

As shown in Tab. 1, the MSCTD contains totally 17,841 bilingual conversations and 142,871/30,370

---

[5]For English↔German, we firstly sample a small set of training data and apply the same test and validation set with the English↔Chinese version. Then, the German translations are collected from professional English-German workers con-

tracted via a language service company (magicdatatech). The three crwodworkers are asked to translate them according to the English utterance, the corresponding image, and the dialogue history.

Dataset	Language Direction	Modality	Scene	Sentiment	#Dialogues			#Instances/Utterances		
					Train	Valid	Test	Train	Valid	Test
Multi30K (Elliott et al., 2016)	English→German/French	T,V	Caption	✗	-	-	-	29,000	1,014	1,000
BSD-AMI-ON (Rikters et al., 2020)	English↔Japanese	T	Dialogue	✗	2,643	69	69	84,800	2,058	2,104
BconTrasT (Farajian et al., 2020)	English↔German	T	Dialogue	✗	550	78	78	13,845	1,902	2,100
BMELD (Liang et al., 2021a)	English↔Chinese	T	Dialogue	✗	1,036	108	274	9,987	1,084	2,601
MSCTD (Ours)	English↔Chinese	T,V	Dialogue	✓	13,749	504	509	132,741	5,063	5,067
	English↔German	T,V	Dialogue	✓	2,066	504	509	20,240	5,063	5,067

Table 2: Comparison of (1) previous mulitmodal machine translation dataset: Multi30k, (2) textual chat translation datasets: BconTrasT, BSD-AMI-ON, and BMELD, and (3) our MSCTD. T/V: text/vision modality.

English-Chinese/English-German utterance pairs with two modalities (*i.e.*, text and image), where each utterance has been annotated with onesentiment label. For English-Chinese/English-German, we split the dialogues into 13,749/2,066 for train, 504/504 for valid, and 509/509 for test while keeping roughly the same distribution of the utterance pair/image, respectively. The detailed annotation of sentiment labels are also listed in Tab. 1, where three labels account for similar proportion.

Based on the statistics in Tab. 1, the average number of turns per dialogue is about 10, and the average numbers of tokens per turn are 8.2, 10.9, and 8.3 for English utterances (word level), Chinese utterances (character level), and German utterance (word level), respectively.

3.5 Related Datasets

The related datasets mainly involve three research fields: multimodal machine translation, textual chat translation, and multimodal dialogue sentiment analysis.

In **multimodal machine translation**, there exists one dataset: Multi30K (Elliott et al., 2016), where each image is paired with one English caption and two human translations into German and French. It is an extension of the original English description dataset: Flickr30K (Young et al., 2014). Afterwards, some small-scale multimodal test sets (about 3k instances) are released to evaluate the system, such as WMT18 test set (1,071 instances) (Barrault et al., 2018).

In **textual chat translation**, three datasets have been released: BSD-AMI-ON (Rikters et al., 2020), BconTrasT (Farajian et al., 2020), and BMELD (Liang et al., 2021a). The BSD-AMI-ON is a document-aligned Japanese-English conversation corpus, which contains three sub-corpora: Business Scene Dialogue (BSD (Rikters et al., 2019)), Japanese translation of AMI meeting corpus (AMI (McCowan et al., 2005)), and Japanese

translation of OntoNotes 5.0 (ON (Marcus et al.)). The BconTrast and BMELD are two human-annotated datasets, which are extended from monolingual textual dialogue datasets Taskmaster-1 (Byrne et al., 2019) and MELD (Poria et al., 2019), respectively.

In **multimodal dialogue sentiment analysis**, the MELD (Poria et al., 2019) and MEISD (Firdaus et al., 2020) datasets are publicly available. The MELD dataset is constructed by extending the EmotionLines (Hsu et al., 2018) from the scripts of the popular sitcom *Friends*. It is similar to MEISD, which is also built from famous English TV shows under different genres (*e.g.*, *Friends*, *Grey's Anatomy*, *The Big Bang Theory*).

The resources mentioned above are extensively used in corresponding fields of research and they even cover some sub-tasks in MSCTD. However, our MSCTD is different from them in terms of both complexity and quantity.

Firstly, multimodal machine translation datasets and textual chat translation datasets are either in multimodal or textual dialogue, while ours includes both. It is obvious that conducting multimodal machine translation in conversations is more challenging due to the more complex scene. Furthermore, MSCTD covers four language directions and contains more than 17k human-annotated utterances-image triplets, which is more than the sum of the annotated ones in Multi30K, BSD-AMI-ON, BconTrasT, and BMELD. Tab. 2 provides information on the number of available modality, dialogues, and their constituent utterances for all the five datasets. What is more, our MSCTD is also annotated with sentiment labels while they are not.

Secondly, compared with two existing multimodal dialogue sentiment analysis datasets, MSCTD's quantity of English version is nearly ten-times of the annotated utterances in MEISD or MELD. More importantly, our MSCTD provides an equivalent Chinese multimodal dialogue senti-

Dataset	#Dialogues			#Utterances		
	Train	Valid	Test	Train	Valid	Test
MELD (Poria et al., 2019)	1,039	114	280	9,989	1,109	2,610
MEISD (Firdaus et al., 2020)	702	93	205	14,040	1,860	4,100
MSCTD-Zh (Chinese version)	13,749	504	509	132,741	5,063	5,067
MSCTD-En (English version)	13,749	504	509	132,741	5,063	5,067
MSCTD-De (German version)	2,066	504	509	20,240	5,063	5,067

Table 3: Comparisons of four multimodal dialogue sentiment analysis datasets: MELD, MEISD, and our MSCTD on two languages.

ment analysis dataset and a relatively small German counterpart. Tab. 3 shows the comparison for all the five datasets, *i.e.*, MELD, MEISD, and our MSCTD on three languages.

4 Image Features

Following previous work (Wang et al., 2018; Ive et al., 2019; Meng et al., 2021), we focus on two types of image representation, namely the coarse-grained spatial visual feature maps and the fine-grained object-based visual features.

Coarse-grained Spatial Visual (CSV) Features.
We use the ResNet-50 model (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) to extract a high-dimensional feature vector $\mathbf{f}_j \in \mathbb{R}^{d_c}$ for image Z_j. These features contain output activations for various filters while preserving spatial information. We refer to models that use such features as **CSV**.

Fine-grained Object-based Visual (FOV) Features. Since using coarse-grained image features may be insufficient to model fine-grained visual elements in images including the specific locations, objects, and facial expressions, we use a bag-of-objects representation where the objects are obtained using an off-shelf Faster R-CNNs (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017). Specifically, for an input image Z_j, we obtain a set of detected objects from Faster R-CNNs, *i.e.*, $\mathbf{O}_j = \{\mathbf{o}_{j,1}, \mathbf{o}_{j,2}, \mathbf{o}_{j,3}, ..., \mathbf{o}_{j,m}\}$, where m is the number of extracted objects and $\mathbf{o}_{j,*} \in \mathbb{R}^{d_f}$. Each object is captured by a dense feature representation, which can be mapped back to a bounding box / region (*i.e.*, Region-of-Interest (ROI)). We refer to models that use such features as **FOV**.

Both types of features have been used in various vision and language tasks such as multimodal dialogue sentiment analysis (Firdaus et al., 2020), image captioning (Xu et al., 2015; Shi et al., 2021), and multimodal machine translation (Ive et al., 2019; Lin et al., 2020; Su et al., 2021).

5 Baseline Models

To provide convincing benchmarks for the MSCTD, we perform experiments with multiple Transformer-based (Vaswani et al., 2017) models for the multimodal chat translation task. Additionally, we provide several baselines for the multimodal dialogue sentiment analysis task.

5.1 Multimodal Chat Translation

According to different visual features, we divide the baselines into three categories: text only (**T**), text plus coarse visual features (**T + CSV**), and text plus fine-grained visual features (**T + FOV**).

T: Trans. (Vaswani et al., 2017): the standard transformer model, which is a sentence-level neural machine translation (NMT) model (Yan et al., 2020; Meng and Zhang, 2019; Zhang et al., 2019), *i.e.*, regardless of the dialogue history. TCT (Ma et al., 2020): A unified document-level NMT model based on Transformer by sharing the first encoder layer to incorporate the dialogue history, which is used as the Textual Chat Translation (TCT) model by (Liang et al., 2021c). CA-TCT (Liang et al., 2021c): A multi-task learning model that uses several auxiliary tasks to help model generate coherence-aware translations.

T+CSV: Trans.+Emb (Vaswani et al., 2017): it concatenates the image feature to the word embedding and then trains the sentence-level NMT model. Trans.+Sum (Ive et al., 2019): it adds the projected image feature to each position of the encoder output. Trans.+Att (Ive et al., 2019): this model utilizes an additional cross-attention sub-layer to attend the image features in each decoder block. MCT: we implement the multimodal self-attention (Yao and Wan, 2020) in the encoder to incorporate the image features into the chat translation model. CA-MCT: similarly, we incorporate image features into the multitask-based chat translation model (Liang et al., 2021c) by the multimodal self-attention.

T+FOV: Trans.+Con (Vaswani et al., 2017): it concatenates the word sequence to the extracted object sequence and thus obtains a new sequence taken as the input of the sentence-level NMT model. Trans.+Obj (Ive et al., 2019): it is a translate-and-refine model (two-stage decoder) where the images are only used by a second-pass decoder. M-Trans. (Yao and Wan, 2020): it leverages a multimodal self-attention layer to encode multimodal information where the hidden representation of im-

Modality	M#	Model	Chinese→English			English→Chinese			German→English			English→German		
			BLEU↑	METEOR↑	TER↓	BLEU↑	METEOR↑	TER↓	BLEU↑	METEOR↑	TER↓	BLEU↑	METEOR↑	TER↓
T	M1	Trans.	19.98	23.46	61.55	24.66	25.25	60.39	21.74	27.87	57.27	21.46	22.91	60.50
	M2	TCT	20.39	24.28	61.32	25.21	25.79	60.17	21.99	27.98	57.71	21.77	23.22	60.35
	M3	CA-TCT	20.83	24.67	60.84	25.62	26.05	59.37	22.27	28.03	56.82	22.19	23.71	59.58
T+CSV	M4	Trans.+Emb*	21.03	24.44	60.54	25.51	26.04	59.79	21.94	27.94	56.70	22.54	23.04	58.96
	M5	Trans.+Sum*	21.29	25.06	60.43	26.06	26.33	58.57	21.99	27.98	56.56	22.02	23.08	59.51
	M6	Trans.+Att*	21.54	25.24	60.35	26.10	26.48	58.29	23.00	28.53	56.52	22.72	23.51	58.04
	M7	MCT (Ours)	22.00	25.46	59.85	26.54	26.75	58.07	23.34	28.71	56.33	23.12	23.94	58.57
	M8	CA-MCT (Ours)	22.51†‡	25.50†	59.34†‡	26.83†‡	26.97†‡	57.72†‡	23.81†‡	28.94‡	55.67†‡	23.48‡	24.21‡	58.33
T+FOV	M9	Trans.+Con*	21.53	24.87	59.56	25.47	26.18	59.00	22.17	28.26	56.02	22.19	23.25	58.07
	M10	Trans.+Obj*	21.82	25.35	59.99	26.24	26.42	57.92	22.41	28.73	55.42	22.88	23.64	57.46
	M11	M-Trans.*	22.38	25.77	59.15	26.60	26.65	57.84	23.40	29.10	55.21	23.18	24.00	57.71
	M12	MCT (Ours)	22.46	25.88	59.27	26.74	26.83	57.59	23.94	29.19	55.03	23.79	24.16	57.65
	M13	CA-MCT (Ours)	22.87†‡	25.94†	58.57†‡	27.04†‡	27.12†‡	57.56†‡	24.33†‡	29.42†	54.90†	24.12†‡	24.41†	57.24†

Table 4: Test results of multimodal chat translation task in terms of BLEU, METEOR, and TER on our MSCTD. The best and the second results are **bold** and <u>underlined</u>, respectively. The symbol '*' denotes sentence-level multimodal machine translation models which do not use the dialogue history. '†' indicates that statistically significant better than the M3 model with t-test $p < 0.01$. '‡' indicates that statistically significant better than the sentence-level multimodal machine translation models (*i.e.*, M13 vs. M11 and M8 vs. M6) with t-test $p < 0.05$.

ages are induced from the text under the guidance of image-aware attention. MCT: here, we incorporate the object-level features into the model instead of coarse one. CA-MCT: similarly, we incorporate the object-level features into the multi-task learning model.

5.2 Multimodal Dialogue Sentiment Analysis

We perform several experiments with different models. text-CNN (Kim, 2014): it only applies CNNs to extract textual information for each utterance in a dialogue. In this approach, we do not use the dialogue history or the additional visual information. DialogueRNN (Majumder et al., 2019): this baseline is a powerful approach for capturing dialogue history with effective mechanisms for sentiment analysis. DialogueRNN + BERT (Firdaus et al., 2020): this model improves the performance of DialogueRNN by using BERT (Devlin et al., 2019) embeddings instead of Glove (Pennington et al., 2014) embeddings to represent the textual features. DialogueRNN + PLM: we propose a stronger baseline built upon the DialogueRNN for sentiment analysis. Specifically, we utilize RoBERTa (Liu et al., 2019) embeddings for English sentiment analysis, and ERNIE (Sun et al., 2019) embeddings for Chinese sentiment analysis, and XLM-R (Conneau et al., 2020) embeddings for German sentiment analysis.

Following Firdaus et al. (2020), we only use the coarse-grained image features (*i.e.*, **CSV**) when training above models with the visual information.

6 Experiments

6.1 Setup

For multimodal chat translation, we utilize the standard Transformer-Base architecture (Vaswani et al., 2017). Generally, we use the settings described in previous work (Ive et al., 2019; Yao and Wan, 2020; Liang et al., 2021c) to conduct experiments on our MSCTD.

For multimodal dialogue sentiment analysis, we mainly follow the settings of previous work (Poria et al., 2019; Firdaus et al., 2020).

Please refer to Appendix A for more details.

6.2 Metrics

For multimodal chat translation, following previous work (Liang et al., 2021c; Ive et al., 2019), we use the SacreBLEU[6] (Post, 2018), METEOR (Denkowski and Lavie, 2014) and TER (Snover et al., 2006) with the statistical significance test (Koehn, 2004) for fair comparison. Specifically, for Chinese→English, we report case-insensitive score. For English→Chinese, the reported score is at the character level. For English↔German, we report case-sensitive BLEU score.

For multimodal dialogue sentiment analysis, following Poria et al. (2019), we report weighted-average F-score.

6.3 Results of Multimodal Chat Translation

Results on English↔Chinese. (1) Among all only text-based models (M1~M3), we find that M1

---

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

Model	Chinese→English		
	BLEU↑	METEOR↑	TER↓
Transformer (**T**)	20.43	24.06	61.00
TCT (**T**)	20.81	24.45	61.19
CA-TCT (**T**)	21.23	24.82	60.75
MCT (**T+CSV**)	22.25	25.60	59.69
CA-MCT (**T+CSV**)	22.68	25.60	59.14

Table 5: Sentiment-aware translation results using ground truth.

Model	Chinese→English			
	BLEU↑	METEOR↑	TER↓	ACC.↑
Transformer (**T**)	20.34	24.01	61.09	64.17
TCT (**T**)	20.78	24.39	60.87	64.65
CA-TCT (**T**)	21.15	24.73	60.74	64.78
MCT (**T+CSV**)	22.31	25.42	59.88	65.26
CA-MCT (**T+CSV**)	22.57	25.51	59.45	65.33

Table 6: Sentiment-aware translation results using predicted sentiment labels. The last column (*i.e.*, ACC.) is the corresponding predicted sentiment accuracy.

performs worse than M2, showing that the dialogue history indeed is beneficial for better translations. Furthermore, M3 can further improve the translation performance, which suggests that modeling the coherence characteristic in conversations is crucial for higher results. These can also be found in other settings (*e.g.*, M7 vs. M4∼6; M8 vs. M7). (2) The models with image features incorporated get higher results than corresponding text-based models (*i.e.*, M4∼M6&M9 vs. M1; M7&M12 vs. M2; M8&M13 vs. M3). (3) The dialogue history and the image features obtain significant cumulative benefits (M8 vs. M1 and M13 vs. M1) (4) Among these image-based models (M4∼M8 or M9∼M13), we observe that different fusion manners of text and image features reflect great difference on effects. It shows that there is much room for further improvement using other more advanced fusion methods. (5) Using **FOV** image features is generally better than the coarse counterpart **CSV** (M9∼M13 vs. M4∼M8), which demonstrates that the fine-grained object elements may offer more specific and effective information for better translations.

Results on English↔German. Similar findings are found on English↔German. This shows that our conclusions are solid and convincing on general datasets. All these results prove the value of our constructed MSCTD.

Furthermore, we provide some stronger baselines that we firstly train the model on the general-domain corpus and then fine-tune it on our chat translation dataset. The results are presented in Table Tab. 8 of Appendix B, which show similar findings observed in Table Tab. 4.

6.4 Effect of Sentiment on Multimodal Chat Translation

To evaluate the effect of sentiment, we conduct some experiments on several baselines including single-modality ones and double-modality ones. In terms of implementation, following Si et al. (2019), we append the sentiment label to the head of the source utterance. Tab. 5 shows the results. Comparing them with the results (M1∼M3 and M7∼M8) without using the sentiment in Tab. 4, we find that using the ground-truth sentiment label has a positive impact on the translation performance. Therefore, we believe that it is a topic worthy of research in the future.

We also conducted the experiments with automatically predicted sentiment labels rather than the gold ones as the reviewer suggested, where we used the mixed sentiment presentation by dot-multiplying the predicted sentiment distribution and the sentiment label representation. The results are shown in Tab. 6, where we find that the sentiment factor, as the inherent property of conversations, indeed has a positive impact on translation performance. We also observe that using the automatically predicted sentiment labels (actually the mixed sentiment representation) shows slightly lower results than using ground truth in terms of three metrics. The reason may be that the mixed sentiment representation has certain fault tolerance.

6.5 Results of Multimodal Dialogue Sentiment Analysis

In Tab. 7, we report the results of sentiment classification on three datasets under different settings.

Results on MSCTD-Zh. We can see that the text-based models perform much poorer than other multimodal systems, which shows that it is not enough to evaluate the sentiment based only on the text. It indicates that visual information and contextual embeddings are crucial for classifying sentiment polarities. Overall, we achieve weighted F1 score of 67.57% with the "DialogueRNN+ERNIE" model.

Results on MSCTD-En/MSCTD-De. On English/German, we observe the same findings on

Model	Modality		MSCTD-Zh (Chinese)				MSCTD-En (English)				MSCTD-De (German)			
	T	V	Pos.	Neu.	Neg.	W-avg.	Pos.	Neu.	Neg.	W-avg.	Pos.	Neu.	Neg.	W-avg.
text-CNN	✓	-	52.69	66.80	60.49	61.19	52.35	63.33	56.12	58.23	42.49	43.06	55.55	48.21
DialogueRNN	✓	-	52.16	69.01	58.77	61.45	52.40	68.10	55.91	60.21	36.38	39.52	62.88	48.50
	-	✓	20.30	37.61	29.11	28.49	20.30	37.61	29.11	28.49	20.30	37.61	29.11	28.49
	✓	✓	51.83	70.21	60.76	62.59	55.51	69.34	57.81	61.69	41.88	46.67	55.36	49.15
DialogueRNN+BERT	✓	-	56.07	73.64	63.39	65.57	58.15	71.90	60.44	64.40	43.33	44.59	61.97	52.76
	✓	✓	57.38	73.73	65.73	66.12	59.15	72.79	61.63	64.99	43.14	46.40	61.87	53.32
DialogueRNN+PLM* (Ours)	✓	-	58.85	73.86	67.92	67.18	59.55	73.30	61.20	65.94	43.49	48.17	62.14	54.19
	✓	✓	60.21	74.13	66.59	**67.57**	58.94	74.27	62.82	**66.45**	44.36	48.14	62.51	**54.46**

Table 7: Test results of multimodal dialogue sentiment analysis task in terms of weighted F-score (%). The "W-avg." denotes weighted-average F-score and the best "W-avg." results are **bold**. The symbol '*' denotes that we use pre-trained language models ERNIE, RoBERTa and XLM-R for Chinese, English, and German, respectively.

Chinese. These show that it is beneficial to introduce the visual information and contextual embeddings into the multimodal dialogue sentiment analysis task for different languages. Overall, we achieve the best F1 score of 66.45% and 54.46% on English and German, respectively.

On this task, we obtain consistent results with previous work (Poria et al., 2019; Firdaus et al., 2020), which suggests the utility and reliability of our MSCTD. Additionally, MSCTD-Zh and MSCTD-De bridge the gap on multimodal dialogue sentiment analysis of Chinese and German.

7 Conclusion and Future Work

In this paper, we introduce a new multimodal machine translation task in conversations. Then, we construct a multimodal sentiment chat translation dataset named MSCTD. Finally, we establish multiple baseline systems and demonstrate the importance of dialogue history and multimodal information for MCT task. Additionally, we conduct multimodal dialogue sentiment analysis task on three languages of the MSCTD to show its added value.

MCT is a challenging task due to the complex scene in the MSCTD, leaving much room for further improvements. This work mainly focuses on introducing the new task and dataset, and we provide multiple models to benchmark the task. In the future, the following issues may be worth exploring to promote the performance of MCT:

- How to effectively perceive and understand the visual scenes to better assist multimodal machine translation in conversations?
- How to build a multimodal conversation representation model to effectively align, interact, and fuse the information of two modalities?

8 Ethical Considerations

In this section, we discuss the main ethical considerations of MSCTD: (1) Intellectual property protection. The English utterance and image of MSCTD is from OpenViDial dataset (Meng et al., 2021). For our translation and sentiments, its permissions are granted to copy, distribute and modify the contents under the terms of the Creative Commons AttributionShareAlike 3.0 Unported License and Creative Commons CC0 License, respectively. (2) Privacy. The data source are publicly available movies. Its collection and Chinese/German annotation procedure is designed for chat translation purpose, and does not involve privacy issues. (3) Compensation. During the sentiment annotation, Chinese and German translation, the salary for annotating each utterance is determined by the average time of annotation and local labor compensation standard. (4) Data characteristics. We refer readers to the content and Meng et al. (2021) for more detailed characteristics. (5) Potential problems. While principled measures are taken to ensure the quality of the dataset, there might still be potential problems with the dataset quality, which may lead to incorrect translations in applications. However, moderate noise is common in large-scale modern translators, even for human translated sentences, which should not cause serious issues.

Acknowledgements

References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of WMT*, pages 304–323.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of EMNLP-IJCNLP*, pages 4516–4525.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of ACL*, pages 992–1003.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of ACL*, pages 1913–1924.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of WMT*, pages 376–380.

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.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

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.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of COLING*, pages 4441–4453.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of LREC*.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of WMT*, pages 639–645.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of ACL*, pages 6525–6538.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pages 1746–1751.

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.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *Proceedings of IJCV*, pages 32–73.

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, 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.

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

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of ACM MM*, page 1320–1329.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of AAAI*, volume 33, pages 6818–6825.

Ralph Weischedel Eduard Hovy Mitchell Marcus, Martha Palmer, Robert Belvin Sameer Pradhan Lance Ramshaw, and Nianwen Xue. Ontonotes: A large training corpus for enhanced processing.

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.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer.

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

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2021. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

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.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of NIPS*, volume 28.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the business conversation corpus. In *Proceedings of WMT*, pages 54–61.

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

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of ACL*, pages 228–234.

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

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing descriptive image captioning with natural language inference. In *Proceedings of ACL*, pages 269–277.

Chenglei Si, Kui Wu, Ai Ti Aw, and Min-Yen Kan. 2019. Sentiment aware neural machine translation. In *Proceedings of WAT*, pages 200–206.

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*.

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, 554:47–60.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

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.

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.

Josiah Wang, Pranava Swaroop Madhyastha, and Lucia Specia. 2018. Object counts! bringing explicit detections back into image captioning. In *Proceedings of NAACL-HLT*, pages 2180–2193.

Longyue Wang, Jinhua Du, Liangyou Li, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. 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, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016. Automatic construction of discourse corpora for dialogue translation. In *Proceedings of the LREC*, pages 2748–2754.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICLR*, pages 2048–2057.

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

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of ACL*, pages 4346–4350.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of ACL*, pages 3025–3035.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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   Implementation Details

For multimodal chat translation, we utilize the standard Transformer-Base architecture (Vaswani et al., 2017). Generally, we use the settings described in previous work (Ive et al., 2019; Yao and Wan, 2020; Liang et al., 2021c) to conduct experiments on our MSCTD. Specifically, we use 512 as hidden size, 2048 as filter size and 8 heads in multihead attention. Both the encoder and the decoder of all the models have 6 layers and are trained using THUMT (Tan et al., 2020). We set the training step to 100,000 steps. The dropout is set to 0.1. The batch size for each GPU is set to 4096 tokens. The experiments are conducted using 4 NVIDIA Tesla V100 GPUs, which gives us about 4*4096 tokens per update. The 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. Label smoothing is set to 0.1. Following Liang et al.

(2021c), we set the number of dialogue context to 3. During inference, the beam size is set to 4, and the length penalty is 0.6 in all experiments.

For the pre-training-then-fine-tuning setting, we firstly train our model on the WMT20 datasets for 100,000 steps. Then, we utilize the pre-trained model to initialize our all multimodal chat translation models.

For multimodal dialogue sentiment analysis, we mainly follow the settings of previous work (Poria et al., 2019; Firdaus et al., 2020). The experiments are conducted on an NVIDIA Tesla V100 GPU and the batch size is set to 64. The learning rate is set to 0.001.

B   Pre-training-then-fine-tuning Results

In this section, we provide some stronger baselines that we firstly train the standard transformer (Vaswani et al., 2017) model on the general-domain corpus (WMT20 dataset of Appendix C) and then fine-tune it on our chat translation dataset (*i.e.*, using the pre-training-then-fine-tuning paradigm). In Tab. 8, the M1 denotes we directly evaluate the pre-trained model on the target chat test set (*i.e.*, without fine-tuning on chat translation dataset.). The M2∼M7 apply the pre-training-then-fine-tuning paradigm. From Tab. 8, we observe similar conclusions to § 6.3. This shows that our findings on the newly proposed dataset are solid even under the stronger baselines. Besides, we also find that, after pre-training on the general-domain corpus, the model obtains significant improvement (M2∼M7 vs. M1).

C   WMT20 Dataset

For English↔Chinese, we combine News Commentary v15, Wiki Titles v2, UN Parallel Corpus V1.0, CCMT Corpus, and WikiMatrix. For English↔German, 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 English↔Chinese and 45,541,367 sentence pairs

Modality	M#	Model	Chinese→English			English→Chinese			German→English			English→German		
			BLEU↑	METEOR↑	TER↓	BLEU↑	METEOR↑	TER↓	BLEU↑	METEOR↑	TER↓	BLEU↑	METEOR↑	TER↓
T	M1	Trans. w/o FT	19.18	24.81	60.72	25.79	27.04	61.24	46.80	38.99	35.19	44.74	35.42	35.48
	M2	Trans.	27.92	29.31	53.92	33.30	29.79	50.92	50.57	44.68	30.21	51.76	39.11	29.96
	M3	TCT	28.28	29.96	51.42	33.94	30.57	50.85	50.80	45.10	30.02	51.96	39.29	29.75
	M4	CA-TCT	<u>28.56</u>	<u>30.24</u>	51.29	<u>34.42</u>	<u>31.11</u>	<u>50.21</u>	51.27	45.21	29.88	52.24	39.36	<u>29.52</u>
T+FOV	M5	Trans.+Con*	28.16	29.88	51.65	33.43	30.10	50.71	51.52	45.87	29.33	52.18	39.45	29.67
	M6	MCT (Ours)	28.49	30.11	<u>51.28</u>	34.07	30.74	50.63	<u>51.75</u>	<u>45.91</u>	<u>29.19</u>	<u>52.45</u>	<u>39.66</u>	29.55
	M7	CA-MCT (Ours)	**28.81**†‡	**30.45**†‡	**51.06**†‡	**34.77**†‡	**31.40**†‡	**50.05**†‡	**51.98**†‡	**46.37**†‡	**29.02**†‡	**52.72**†‡	**39.58**†	**29.39**†

Table 8: Pre-training-then-fine-tuning results of multimodal chat translation task in terms of BLEU, METEOR, and TER on Test set. The best and the second results are **bold** and underlined, respectively. The symbol '*' denotes sentence-level multimodal machine translation models which do not use the dialogue history. '†' and '‡' indicates that statistically significant better than the M2 model with t-test $p < 0.01$ and $p < 0.05$, respectively.

for English↔German, respectively.