Context-aware Neural Machine Translation for English-Japanese Business Scene Dialogues

Sumire Honda   pu.sumirehonda@gmail.com
Computational Linguistics, University of Potsdam, Potsdam, Germany
Patrick Fernandes   pfernand@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal
Chrysoula Zerva   chrysoula.zerva@tecnico.ulisboa.pt
Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal

Abstract

Despite the remarkable advancements in machine translation, the current sentence-level paradigm faces challenges when dealing with highly-contextual languages like Japanese. In this paper, we explore how *context-awareness* can improve the performance of the current Neural Machine Translation (NMT) models for *English-Japanese business dialogues* translation, and what kind of context provides meaningful information to improve translation. As business dialogue involves complex discourse phenomena but offers scarce training resources, we adapted a pretrained mBART model, finetuning on multi-sentence dialogue data, which allows us to experiment with different contexts. We investigate the impact of larger context sizes and propose novel context tokens encoding extra-sentential information, such as speaker turn and scene type. We make use of *Conditional Cross-Mutual Information* (CXMI) to explore how much of the context the model uses and generalise CXMI to study the impact of the *extra-sentential context*. Overall, we find that models leverage both preceding sentences and extra-sentential context (with CXMI increasing with context size) and we provide a more focused analysis on honorifics translation. Regarding translation quality, increased source-side context paired with scene and speaker information improves the model performance compared to previous work and our context-agnostic baselines, measured in BLEU and COMET metrics. [1]

1 Introduction

Traditionally NMT models such as Transformers (Maruf et al., 2021) approach the task of machine translation (MT) focusing on individual sentences without considering the surrounding information, such as previous utterances or underlying topics. As a result, the output often lacks discourse coherence and cohesion, which is problematic for MT applications such as chat translation systems (Farajian et al., 2020; Bawden et al., 2018). Thus, it is still an open research question to what degree these models can take advantage of contextual information to produce more accurate translations.

To answer this question, several context-aware NMT (Tiedemann and Scherrer, 2017; Voita et al., 2019; Maruf et al., 2019; Xu et al., 2021) studies have been conducted by adding

---

[1]Code available at: `https://github.com/su0315/discourse_context_mt`

surrounding sentences to the models and testing if it helps to capture better specific linguistic phenomena requiring context (e.g. coreference resolution). However, there is limited work on discourse or dialogue datasets, and most of it is focused on high-resource or Indo-European (IE) languages (Liu et al., 2021). Therefore, there is a need to investigate how well do the proposed approaches capture discourse phenomena in non-IE or low-resource languages.

This work aims to address the aforementioned gap by focusing on English-Japanese (En-Ja) translation for business dialogue scenarios in order to examine if current context-aware NMT models (Tiedemann and Scherrer, 2017) actually use the additional context, and what kind of context is useful regarding the translation of linguistic phenomena pertaining to Japanese discourse, such as honorifics. We specifically propose the use of novel extra-sentential information as additional context and show that it improves translation quality. Overall, the main contributions of this study are threefold: (1) We demonstrate that it is possible to adapt a (non-context-aware) large pretrained model (mBART; Liu et al. (2020); Tang et al. (2021)) to attend to context for business dialogue translation and propose an **improved attention mechanism** (CoAttMask) with significant performance gains for source-side context, even on small datasets; (2) we propose **novel extra-sentential information** elements such as speaker turn and scene type, to be used as additional source-side **context**; and (3) we compare the use of context between our context-aware models using CXMI (Fernandes et al., 2021), a mutual-information-based metric and perform a more focused analysis on the translation of **honorifics**.

2 Related Work

2.1 Context-aware MT

Context-aware MT lies between sentence-level MT and document-level MT, as the former assumes the translation of a single sentence from source to target language with no other accessible content, and the latter implies the translation of a sequence of sentences from a document, assuming access to the whole document. Context-aware MT lies close to the definition of document-level MT, as it requires access to context either in the form of preceding sentences or other type of information regarding the topic and setup of the text to be translated, that can aid in its translation.

Several methods using a transformer-based architecture (Vaswani et al., 2017) have been proposed for context-aware NMT, frequently categorised into single-encoder and multi-encoder models (Sugiyama and Yoshinaga, 2019). Single-encoder models concatenate the source sentence with (a) preceding sentence(s) as the contexts, with a special symbol to distinguish the context and the source or target in an encoder (Tiedemann and Scherrer, 2017). Multi-encoder models pass the preceding sentence(s) used as context through a separate encoder modifying the Transformer architecture (Voita et al., 2018; Tu et al., 2018). According to Sugiyama and Yoshinaga (2019), the observed performance gap between the two models is marginal, but the single-encoder models are relatively simpler architectures without modifying sequence-to-sequence transformers.

Apart from concatenating preceding sentences on the source-side, some works focus on the target-side context, i.e., show some benefits from attempting to decode multiple sequential sentences together (Su et al., 2019; Mino et al., 2020). Depending on the use-case, source-side, target-side, or a combination of contexts has proven beneficial (Agrawal et al., 2018; Chen et al., 2021; Fernandes et al., 2021). Additionally, some works focused more on context related to discourse phenomena, with Liang et al. (2021a) proposing the use of variational autoencoders to model dialogue phenomena such as speaker role as latent variables (Liang et al., 2021b). We examine here a simpler approach, that directly encodes such speaker and scene information and allows the model to use it as additional context. In more recent work, the impact of pretraining on larger out-of-domain (OOD) data has also been studied to aid in downstream MT tasks with limited resources (Voita et al., 2019; Liang et al., 2022).

For English-Japanese translation, there have been some context-aware NMT studies that used variations of single-encoder models in the news and dialogue domain (Sugiyama and Yoshinaga, 2019; Ri et al., 2021; Rikters et al., 2020). Specifically for dialogue, Rikters et al. (2020) experimented with context-aware MT that employs source-side factors on Ja-En (Japanese-English) and En-Ja (English-Japanese) discourse datasets. They propose to concatenate the preceding sentence(s) from the same document followed by a tag-token to separate the context from the original sentence and use binary token-level factors on top of this to signify whether a token belongs to the context or source sentence.

2.2 Japanese Honorifics in NMT

For into-Japanese MT, specific discourse phenomena such as honorifics constitute a core challenge when translating from languages that do not include such phenomena, like English (Hwang et al., 2021; Sennrich et al., 2016). Japanese honorifics differ to English because different levels of honorific speech are used to convey respect, deference, humility, formality, and social distance, using different types of verbal inflexions. Besides, the desired formality is decided depending on social status and context and may involve more extensive changes in utterances compared to other languages (Fukada and Asato, 2004). Feely et al. (2019) proposed formality-aware NMT, conditioning the model on a manually selected formality level to evaluate honorifics. They evaluate the formality level of the translated sentences using their formality classifier, showing improvements. Instead of explicitly selecting the formality level, we evaluate the impact of our context representations on the correct translation of honorifics, inspired by Fernandes et al. (2023).

3 Datasets

We use Business Scene Dialogue corpus (BSD) (Rikters et al., 2019) as the main dataset. Additionally, only to compare the performance in a certain setup with the main dataset, we also use AMI Meeting Parallel Corpus (AMI) (Rikters et al., 2020) as a supplemental dataset. They are both document-level parallel corpora consisting of different scenes (dialogue sequence scenarios) or meetings and include both out-of-English and into-English translations, of which we use the English-Japanese translation direction. We focus our analysis on the BSD dataset, as it contains more scenarios and extra-sentential information which we use as additional context.

In the main dataset BSD, each document consists of a business scene with a scene tag (face-to-face, phone call, general chatting, meeting, training, and presentation), and each sentence has speaker information that indicates who is speaking. Contents of BSD are originally written either in English or Japanese by bilingual scenario writers who are familiar with business scene conversations and then translated into the other language to create a parallel corpus.

As for AMI, the contents are translations to Japanese from 100 hours of meeting recordings in English. Since it originates from naturally occurring dialogue it contains shorter utterances than BSD, including multiple single-word sentences with filler and interjection words. The data split statistics for BSD and AMI are shown in Table 1. The domain of BSD and AMI is similar, however, AMI does not include scene information and the number of documents (scenarios) is smaller.

	BSD	Train	Dev	Test	**AMI**	Train	Dev	Test
Sentences		20,000	2051	2120		20,000	2000	2000
Scenarios		670	69	69		30	5	5

Table 1: Data split statistics for BSD and AMI dataset

model	source side	target side	model
1-1	First of all, I want to thank you for all your hard work.	君の勤勉さにはとても感謝している という事をまず最初に伝えたい。	1-1
	+context size 2	*+context size 2*	
3-1	Mr Billy, I was told you needed me in your office. **</t>** Thank you for coming. **</t>** First of all, I want to thank you for all your hard work.	ビリーさん、用があると聞いたのですが。**</t>** 来てくれてありがとう。**</t>** 君の勤勉さ にはとても感謝しているという事をまず 最初に伝えたい。	1-3
	+speaker tags		
3-1 + speaker	*<SameSpeaker>* Mr Billy, I was told you needed me in your office. **</t>** *<DiffSpeaker>* Thank you for coming. **</t>** *<SameSpeaker>* First of all, I want to thank you for all your hard work.		

Figure 1: Context-extended inputs on source and target side. Coloured text corresponds to added context, **bold** signifies context separators and ***bold-italics*** speaker-related context tags.

4 Methodology

In this section, we analyse our context-aware NMT approach in a dialogue setup in two steps: firstly, we consider what type of information might be useful as context and how it should be encoded to generate useful input representations, and secondly, we discuss modifications in the original encoder-decoder architecture that facilitate learning to attend to context even when tuning on small datasets.

4.1 Encoding Context

We adapt the method of Tiedemann and Scherrer (2017) and experiment with encoding contexts both on source-side and target-side. Unlike Tiedemann and Scherrer (2017) who considers a single preceding sentence, we experiment with up to five preceding sentences, motivated by the findings of Fernandes et al. (2021); Castilho et al. (2020). We intercept a separator token </t> following every context sentence as shown in Figure 1.

We compare the context-aware models to the context-agnostic model, finetuned on our dataset. Henceforth, in this work, we will refer to the context-agnostic model as a 1-1 model, meaning that the model's source-side input is only 1 source sentence, and the target-side input is also only 1 target sentence during the training. For the context-aware models, this paper uses the naming convention of 2-1, 3-1, 4-1, and 5-1 for source context-aware models and 1-2, 1-3, 1-4, and 1-5 for target context-aware models. Note that in this work we use the gold data (human-generated translations of previous sentences) to represent the target context. Although the accessibility of target-side context data is limited in real-world translation tasks, there are some relevant use cases. For example, in a chatbot system where a human can edit the predicted translation in preceding sentences before the current sentence translation, the gold label of preceding target-side sentences is accessible.

Speaker Information: Delving deeper into the dialogue scenario, we also explore whether speaker-related information can provide useful context. In a dialogue dataset with multiple speakers, each speaker may utter a varying number of sentences per turn, and as such using a fixed context window implies potentially including multiple speakers in the context. Since aspects such as discourse style, politeness, honorifics in Japanese (Feely et al., 2019) or even topic distribution can be tied to specific speakers, knowing when a speaker changes in the context can be particularly informative. Speaker information has been used to improve user experience in simultaneous interpretation (Wang et al., 2022), but to the best of our knowledge, it has not been explored as a contextual feature for MT.

Hence, we consider two speaker types: (1) the one who utters the sentence to be translated – and who may have communicated more sentences in the context window – (same speaker) and (2) any other speaker(s) with utterances within the context window (different speaker), between

which we do not differentiate. In other words, we only encode information about whether there has been **a change of speakers** within the context. We achieve this by concatenating either a special token `<DiffSpeak>` (Different speaker) or a `<SameSpeak>` (Same speaker) to each sentence (utterance) of the context as shown in the last row of Figure 1. This example also highlights the potential difference in speaker formality: the boss uses more casual expressions compared to the employee.

Scene Information: Similar to speaker information, we consider the information associated with the dialogue scene and its potential impact on the translation if used as context. We hence experiment with an additional special token representing the scene tag in BSD dataset. Following BSD dataset scene tags explained in §3, we prepared six additional tokens; `<face-to-face conversation>`, `<phone call>`, `<general chatting>`, `<meeting>`, `<training>`, and `<presentation>`. One of the tags is concatenated at the very beginning of each source input to signify the scene of the dialogue. For example, the scene tag of conversation in Figure 1 is `<face-to-face conversation>`, so the 2-1 model's input will be "`<face-to-face conversation>` *Thank you for coming. `</t>` First of all, I want to thank you for all your hard work.*". Such information could provide a useful signal regarding the speaker style, such as honorifics and formality, or even scene-specific terminology.

4.2 Context-aware Model Architecture

To encode context we rely on the Tiedemann and Scherrer (2017) approach, which we adapt to optimise performance for the BSD dataset. Due to the small size of available datasets for the business dialogue scenarios it is difficult to train a context-aware transformer architecture from scratch. Instead, we opt for fine-tuning a multi-lingual large pretrained model.

Baseline: All the models for En-Ja translation in this experiment are finetuned with mBART50 (Liu et al., 2020; Tang et al., 2021) with our proposed architectural modification for context-aware models described in the following paragraphs. We train all models until convergence on the validation set and use a `max_token_length` of size 128 for the baseline model, and 256 for the context-aware ones [2]. mBART is one of the state-of-the-art multilingual NMT models, with a Transformer-based architecture (Vaswani et al., 2017). It follows BART (Lewis et al., 2020) Seq2Seq pretraining scheme and is pretrained in 50 languages, including Japanese and English, using multilingual denoising auto-encoder strategy.

Target context-aware model: To consider context on the target side we essentially decode the target-context as shown in Figure 1 instead of a single sentence. To apply the Tiedemann and Scherrer (2017)'s context-aware approach to the target-side, the baseline model architecture was modified to prevent the loss function from accounting for mispredicted context and optimising instead only for the original target sentence.

Source context-aware model: Contrary to (Tiedemann and Scherrer, 2017; Bawden et al., 2018) we found that directly using the extended source inputs resulted in significantly lower performance for all context sizes, when compared to the original context-agnostic model (see Table 2). We attribute this inconsistency in our findings to the small size of the BSD dataset which might be insufficient for tuning a large pretrained model towards a context-aware setup.

To address this issue, a new architecture Source **Co**ntext **Att**ention **Mask** Model (CoAttMask) is proposed. In this approach, we pass the context-extended input to the encoder part of the model but mask the encoder outputs that correspond to the context when passed to the decoder. As shown in the yellow block in Figure 2, after the context-extended input is passed to the encoder, we mask the context-related part when passing the encoded input to the decoder to compute cross attention. As such, the context is leveraged to compute better input representations through self-attention in the transformer but does not further complicate the

---

[2]All hyperparameters are at: `https://github.com/su0315/discourse_context_mt`

decoding process. Table 2 shows that the CoAttMask model successfully outperformed the baseline model architecture (without CoAttMask).

5 Evaluation

5.1 Metrics for Overall Performance

To report the performance of the MT models, we report BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) scores. We use COMET as the primary metric since it has shown to be more efficient in assessing MT quality, better capturing valid synonyms and paraphrases (Smith et al., 2016) as well as discourse phenomena in longer text (Maruf et al., 2021).

5.2 Metric for Context Usage – CXMI –

Although COMET can capture more semantic features than BLEU, it is still difficult to assess how much context-aware NMT models actually use the additional contexts to improve predictions. To that end, we use Conditional Cross Mutual Information (CXMI) (Bugliarello et al., 2020; Fernandes et al., 2021). CXMI measures the entropy (information gain) of a context-agnostic machine translation model and a context-aware machine translation model. The CXMI formula can be seen in Eq. (1), where C signifies additional context, Y the target, X the source, $H_{q_{MT_A}}$ the entropy of a context-agnostic machine translation model, and $H_{q_{MT_C}}$ the entropy of context-aware machine translation model. Thus, a positive CXMI score indicates a useful contribution of context to predicting the correct target (increasing the predicted score of the correct target words). This can be estimated with Eq. (2), over a test dataset with N sentences, when $y^{(i)}$ is ith target sentence and $x^{(i)}$ the ith source sentence in each document (Fernandes et al., 2021).

$$CXMI(C \rightarrow Y|X) = H_{q_{MT_A}}(Y|X) - H_{q_{MT_C}}(Y|X,C) \quad (1)$$

$$\approx -\frac{1}{N}\sum_{i=1}^{N}\log\frac{q_{MT_A}(y^{(i)}|x^{(i)})}{q_{MT_C}(y^{(i)}|x^{(i)},C^{(i)})} \quad (2)$$

In this experiment, CXMI is calculated between context-aware models with preceding sentence(s), speaker information, and scene information and each corresponding baseline model that lacks the respective context. To compute CXMI, a single model that can be tested with both context-agnostic inputs and context-extended inputs is required. We hence train the models with dynamic context size, such that during training the model can see anywhere from 0 to k context sentences (Fernandes et al., 2021).

5.3 *Honorifics P-CXMI*

To evaluate how much additional context is actually used to improve translation with respect to honorifics, we also compute P-CXMI, an extension of CXMI that allows us to measure the impact

Context Size	Baseline	CoAttMask
0	0.724	-
1	0.661	0.724
2	0.665	0.724
3	0.662	**0.727**
4	0.658	**0.727**

Table 2: Performance of CoAttMask model in COMET. **Bold** scores signify the performance improved 1-1 model
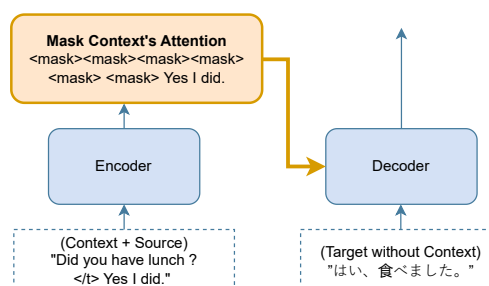

Figure 2: CoAttMask Architecture

of context on specific translations or words in a translation instead of over the whole corpus (Fernandes et al., 2023). We define *Honorifics P-CXMI* for token-level honorific expressions, which we calculate only for cases where the gold label is an honorific expression. While CXMI is calculated on the corpus level, averaged over the number of sentences, *Honorifics P-CXMI* is calculated for each honorific token and averaged over the number of the honorific tokens in the testset. As such, it is not directly comparable to the CXMI values (Fernandes et al., 2023).

Inspired by Japanese honorific word lists proposed in Fernandes et al. (2023) and Farajian et al. (2020), the following tokens are selected as the main honorific expressions (based on frequency of use and non-ambiguous functionality in the sentence) [3] "です (desu)", "でした (deshita)", "ます (masu)", "ました (mashita)", "ません (masen)", "ましょう (mashou)", "でしょう (deshou)", "ください (kudasai)", "ございます (gozaimasu)", "おります (orimasu)", "致します (itashimasu)", "ご覧 (goran)", "なります (narimasu)", "伺 (ukaga)", "頂く (itadaku)", "頂き (itadaki)", "頂いて (itadaite)", "下さい (kudasai)", "申し上げます (moushiagemasu)". Those tokens are mainly categorized as three types of honorifics: respectful (sonkeigo, 尊敬語), humble (kenjogo, 謙譲語), polite (teineigo, 丁寧語).

6 Experimental Results

We compare our work to previous approaches evaluated on BSD, namely this of Rikters et al. (2019) who combined multiple En-Ja datasets to train a model for En-Ja dialogue translation and Rikters et al. (2021) who also used a context-aware variant of Tiedemann and Scherrer (2017) combined with factors to encode dialogue context. Additionally, we compare with our context agnostic baseline. Table 3 shows that tuning mBART on the BSD data already outperformed the previous studies by more than 9 points in terms of BLEU, highlighting the impact of pretraining on large multilingual data. For the context-aware models, four types of models are compared for different context sizes; (1) Preceding Sentences Model (§6.1); (2) Speaker Information Model; 3) Scene Information Model; and (4) Speaker & Scene Information Model (§6.2).

6.1 Context-aware Models: Preceding Sentences

As seen in Table 3, as we increase the size of the context used, the CXMI score consistently increases indicating better leveraging of the context provided for the prediction of the target words. However, this increased attention to context is only reflected in small gains in the overall performance for specific context sizes. Specifically, for the source-side context only the models with larger context of 3 and 4 sentences improved for BLEU and COMET, as opposed to previous work that observes gains on single sentence context and often decreasing performance for larger context sizes (Tiedemann and Scherrer, 2017; Voita et al., 2018; Rikters et al., 2020; Ri et al., 2021; Nagata and Morishita, 2020). We hypothesize that this relates to our stronger baseline, and the specifics of the dialogue translation task: shorter utterances on average and multiple speakers which could lead to useful context lying further away in the dialogue history.

For the target-side context most variants either under-performed or performed similarly to the context-agnostic model. Indeed, while we notice an increased usage of context as we increase the target context size (see Figure 3), this does not seem to lead to improved performance. Further supported by the findings in §6.3 on the AMI dataset, it seems that using context on the source side is more beneficial for such small dialogue datasets and we focus our analysis and experiments more on the source side. However, it would be interesting to consider further adapting target-side context or explore pre-training on larger corpora as a way to mitigate this in future work (Liang et al., 2022; Su et al., 2019).

Focusing on CXMI as shown in Table 3 and Figure 3, our experiments corroborate the main findings of Fernandes et al. (2021). We can see that for both target and source the biggest jump

---

[3]Modified for the mBART50 tokenizer.

	Model (context size)	BLEU ↑	COMET ↑	CXMI ↑
Baselines	Rikters et al. (2019) (0)	13.53	-	-
	Rikters et al. (2021) (0)	12.93	-	-
	Rikters et al. (2021) (1)	14.52	-	-
	Ri et al. (2021) (1)	17.11	-	-
	1-1 (0)	26.04	0.725	0
Source context	2-1 (1)	25.87	0.724	0.32
	3-1 (2)	25.41	0.724	0.36
	4-1 (3)	**26.09**	**0.727**	0.38
	5-1 (4)	**26.09**	**0.727**	0.39
Target context	1-2 (1)	25.85	0.72	0.65
	1-3 (2)	**26.08**	0.702	0.76
	1-4 (3)	25.77	0.704	0.83
	1-5 (4)	24.96	0.71	0.88

Table 3: Score comparison between preceding sentences models and 1-1 model. **Bold** scores signify the performance improved baseline (BLEU, COMET)
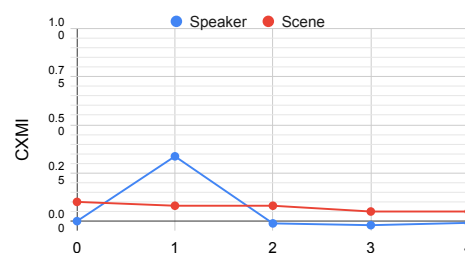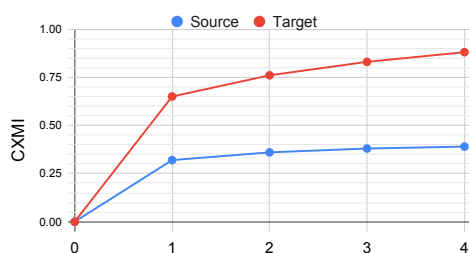

Figure 3: CXMI for source and target context-aware models in each context size

Figure 4: CXMI for speaker and scene model in each context size

in context usage is when we increase the context size from 0 to 1, but unlike Fernandes et al. (2021) we subsequently observe small but consistent increases for each context size (ascending).

Table 4 shows the result of *Honorifics CXMI* between source-side preceding sentences models and 1-1 model. With respect to the translation of honorifics, *Honorifics CXMI* scores for all context sizes show positive score, indicating that the provision of additional context helps the model to attribute higher density to the correct honorific translation. In other words, the model can leverage additional context to improve the prediction of honorific expressions.

Looking at the improved scores for each context size and honorific expression separately, we found that in all cases, it was the translation of the honorific token "伺 (ukaga)" that benefited the most. "伺 (ukaga)" is an honorific token that is a component of "伺う(ukagau)", a verb meaning "go" or "ask" in Japanese honorific expression. In particular, "伺う(ukagau)" is one of the humble (kenjogo, 謙譲語) expressions, and the humble is used in a business email or very formal speech (Liu and Kobayashi, 2022). These honorific expressions are used strictly by speakers to refer to themselves when they address a superior in business settings (Rahayu, 2013). As such, previous utterances that would reveal the relation of the speaker to the addressee are necessary to obtain the correct translation. Table 5 demonstrates the correction in the use of "伺 (ukaga)" when using a context window of size 2. The baseline model predicts "申します" instead of "伺 (ukaga)", leading to a semantically inappropriate translation meaning "*I'm (Takada)*" while with additional context it correctly predicts the "伺 (ukaga)" token.

	2-1	3-1	4-1	5-1
Honorifics CXMI ↑	0.05	0.07	0.06	0.06

Table 4: *Honorifics CXMI* between source-side preceding sentences models and 1-1 model

Source Sentence	Reference Sentence	1-1 Model Prediction	3-1 Model Prediction
I, Takada from Company I will <u>go</u> to your place at 5 o'clock in the afternoon tomorrow.	明日の午後5時に、わたくし、I社の高田が<u>伺います</u>。	明日の午後5時に、I社の高田と申します。	私、I社の高田が明日の午後5時に御社へお<u>伺い</u>します。

Table 5: Comparison between a context-agnostic model (1-1) and a context-aware model (3-1) in predicting honorific token "伺". (<u>Underlined</u> words signify that the 3-1 model improved the 1-1 model in predicting the correct token.)

6.2 Extra-sentential context:

For the following experiments, we focus on further enhancing the source-side context by adding scene and speaker information as discussed in §4.1. We first explore their usefulness separately, concatenating to the context either speaker tags or scene tags, as shown in Table 6 and Figure 4.

Speaker Information Models: When adding speaker information ("With Speaker", Table 6) the model seems to be obtaining slightly better performance on BLEU scores but not COMET. Additionally, with respect to the CXMI (see Figure 4), the speaker information seems to be useful for the model predictions only when using a single sentence of context. In other words, the model benefits only from knowing whether the previous utterance originated from the same speaker or not. While this finding is quite intuitive (a change of speaker could indicate a switch in style and formality) it is still unclear why this does not hold for larger context windows.

Note that while the benefits of using the speaker turn information seem limited, there are further aspects to be explored that were out of scope in this work. Specifically, given sufficient training data one could use a separate tag for each speaker in case of ≤ 2 speakers, either using abstract speaker tags, or even the speaker names, potentially helping toward pronoun translation.

Model (Context Size)	Preceding Sentences		With Speaker		With Scene		With Speaker & Scene	
	BLEU↑	COMET↑	BLEU↑	COMET↑	BLEU↑	COMET↑	BLEU↑	COMET↑
1-1 (0)	26.04	0.725	-	-	**26.19**	**0.726**	-	-
2-1 (1)	25.87	0.724	25.94	0.718	**26.18**	0.727	**26.18**	**0.730**
3-1 (2)	25.41	0.724	26.09	0.722	26.26	0.727	<u>**26.41**</u>	<u>**0.740**</u>
4-1 (3)	26.09	0.727	26.03	0.722	**26.27**	**0.731**	26.07	0.730
5-1 (4)	26.09	0.727	**26.39**	0.726	26.1	**0.728**	26.15	0.720

Table 6: Score comparison among preceding sentence models (w/o speaker and scene information), and models with addition of speaker and scene tags. **Bold** scores signify the best performance for each context size and <u>underlined</u> ones the best performance overall.

Scene Information Model: Unlike the speaker information, scene information can be added when the context size is zero too, since it does not need preceding sentences.

In contrast to speaker information models, "With Scene" models outperformed "Preceding Sentences" models for both BLEU and COMET on all context sizes, including when used with no additional context. Additionally, CXMI remains positive for all context sizes with a small decrease when the context size is larger. Hence, we can conclude that scene information helps

towards the correct translation especially when limited context is available.

Speaker and Scene Model: We finally investigate if combining scene and speaker information can further improve performance. Indeed, for smaller context windows (speaker & scene models 2-1 and 3-1) outperformed their respective scene-only and speaker-only versions. Also, the 3-1 speaker & scene model obtained the best performance overall. Hence, while speaker information on its own did not improve performance, the combination of speaker information and scene information outperformed the models without them. This finding indicates that for specific scenarios (scenes), speaker turn might provide more useful signal. Indeed, depending on the scene the speakers may change more or less frequently signifying a necessary change of style (e.g. compare a presentation scene versus the phone call one). It would be interesting to further explore the relationship between the speaker switch frequency and scene type in the future.

6.3 Performance on the AMI dataset

To examine the context-aware model's performance on a similar dataset, we also tested the trained preceding sentences models using AMI dataset introduced in §3. Table 7 shows the performance of the context-aware models on increasing context size. Both context-aware and context-agnostic models obtain higher scores on the AMI dataset, compared to BSD. We notice however that we obtain small performance boosts for some context-aware combinations. More importantly, CXMI findings corroborate those on BSD: as the context size gets larger, CXMI increases both on source and target side. The similar CXMI trends reinforce our findings, hinting that they are not artifacts of a specific dataset, but rather a property of the language pair.

	Baseline	Source Side				Target Side			
	1-1	2-1	3-1	4-1	5-1	1-2	1-3	1-4	1-5
BLEU	32.46	**32.8**	32.12	**32.61**	32.05	32.13	31.22	31.29	**32.56**
COMET	0.852	**0.858**	0.846	**0.854**	0.846	0.848	0.833	0.833	0.85
CXMI	-	0.24	0.27	0.31	0.34	0.07	0.17	0.25	0.48

Table 7: Score comparison between preceding sentences models and 1-1 models with AMI dataset. **Bold** scores signify the performance improved over the baseline (BLEU, COMET).

7 Conclusion and Future Work

This paper explored to what degree encoded context can improve NMT performance for English-Japanese dialogue translation, and what kind of context provides useful information. With our proposed method, we were able to tune mBART on small dialogue datasets and obtain improved MT performance using context. We found that source-side context was more beneficial towards performance and that complementing our source-side context with scene and speaker-turn tags provided further performance improvements. We further analyse the impact of our proposed context-aware methods on the translations obtained, with a focus on translation of Japanese honorifics. In future work, we aim to further investigate context for dialogue translation, expanding to a multilingual setup, larger datasets, and additional extra-sentential context.

Acknowledgements

References

Agrawal, R. R., Turchi, M., and Negri, M. (2018). Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Bugliarello, E., Mielke, S. J., Anastasopoulos, A., Cotterell, R., and Okazaki, N. (2020). It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Castilho, S., Popović, M., and Way, A. (2020). On context span needed for machine translation evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

Chen, L., Li, J., Gong, Z., Duan, X., Chen, B., Luo, W., Zhang, M., and Zhou, G. (2021). Improving context-aware neural machine translation with source-side monolingual documents. In *IJCAI*, pages 3794–3800.

Farajian, M. A., Lopes, A. V., Martins, A. F. T., Maruf, S., and Haffari, G. (2020). Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

Feely, W., Hasler, E., and de Gispert, A. (2019). Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China. Association for Computational Linguistics.

Fernandes, P., Yin, K., Martins, A. F., and Neubig, G. (2023). When does translation require context? a data-driven, multilingual exploration. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Fernandes, P., Yin, K., Neubig, G., and Martins, A. F. T. (2021). Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Fukada, A. and Asato, N. (2004). Universal politeness theory: application to the use of japanese honorifics. *Journal of pragmatics*, 36(11):1991–2002.

Hwang, Y., Kim, Y., and Jung, K. (2021). Context-aware neural machine translation for korean honorific expressions. *Electronics*, 10(13):1589.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liang, Y., Meng, F., Chen, Y., Xu, J., and Zhou, J. (2021a). Modeling bilingual conversational characteristics for neural chat translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5711–5724.

Liang, Y., Meng, F., Xu, J., Chen, Y., and Zhou, J. (2022). Scheduled multi-task learning for neural chat translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4375–4388.

Liang, Y., Zhou, C., Meng, F., Xu, J., Chen, Y., Su, J., and Zhou, J. (2021b). Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 67–79.

Liu, M. and Kobayashi, I. (2022). Construction and validation of a Japanese honorific corpus based on systemic functional linguistics. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.

Liu, S., Sun, Y., and Wang, L. (2021). Recent advances in dialogue machine translation. *Information*, 12(11):484.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Maruf, S., Martins, A. F., and Haffari, G. (2019). Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102.

Maruf, S., Saleh, F., and Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Mino, H., Ito, H., Goto, I., Yamada, I., and Tokunaga, T. (2020). Effective use of target-side context for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4483–4494, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nagata, M. and Morishita, M. (2020). A test set for discourse translation from Japanese to English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3704–3709, Marseille, France. European Language Resources Association.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rahayu, E. T. (2013). The japanese keigo verbal marker. *Advances in Language and Literary Studies*, 4(2):104–111.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ri, R., Nakazawa, T., and Tsuruoka, Y. (2021). Zero-pronoun data augmentation for Japanese-to-English translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 117–123, Online. Association for Computational Linguistics.

Rikters, M., Ri, R., Li, T., and Nakazawa, T. (2019). Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pages 54–61, Hong Kong, China. Association for Computational Linguistics.

Rikters, M., Ri, R., Li, T., and Nakazawa, T. (2020). Document-aligned japanese-english conversation parallel corpus. In *Proceedings of the Fifth Conference on Machine Translation*, pages 637–643, Online. Association for Computational Linguistics.

Rikters, M., Ri, R., Li, T., and Nakazawa, T. (2021). Japanese–english conversation parallel corpus for promoting context-aware machine translation research. *Journal of Natural Language Processing*, 28(2):380–403.

Sennrich, R., Haddow, B., and Birch, A. (2016). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

Smith, A., Hardmeier, C., and Tiedemann, J. (2016). Climbing mont BLEU: The strange world of reachable high-BLEU translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 269–281.

Su, J., Zhang, X., Lin, Q., Qin, Y., Yao, J., and Liu, Y. (2019). Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding. *Artificial Intelligence*, 277:103168.

Sugiyama, A. and Yoshinaga, N. (2019). Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Tu, Z., Liu, Y., Shi, S., and Zhang, T. (2018). Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Wang, X., Utiyama, M., and Sumita, E. (2022). A multimodal simultaneous interpretation prototype: Who said what. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 132–143, Orlando, USA. Association for Machine Translation in the Americas.

Xu, H., Xiong, D., Van Genabith, J., and Liu, Q. (2021). Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3933–3940.