Individual Project Module: Context-aware Neural Machine Translation for English-Japanese Business Scene Dialogues

Sumire Honda[*], **Chrysoula Zerva** (Supervisor)[**], and **David Schlangen** (Supervisor)[*]

[*]University of Potsdam
[**]Instituto Superior Técnico, Instituto de Telecomunicações

Abstract

This paper aims to analyse whether context-enhanced inputs improve the performance of the current Neural Machine Translation (NMT) models in an English-Japanese business scene dialogue dataset, and what kinds of contexts provide meaningful information. The experiment is conducted with the mBART autoencoder, and it is adapted to consider the context (previous utterances) both on the source and target side. Also, the use of novel meta-information elements such as speaker turn and scene type as additional source-side context is proposed.

From the experimental results, we find that increasing the size (number of sentences) of source-side context and the addition of scene information improve the model performance. Overall, the best context-aware model outperformed our context-agnostic baseline by 0.37 in BLEU and 0.015 in COMET score. Furthermore, the CXMI score – a mutual-information-based metric that allows us to compare the use of context between different models, is employed. We find that increasing the context-enhanced information leads to increased CXMI scores, and we carry out *Honorifics CXMI* – an analysis to show that context-enhanced information is used to improve translations of honorifics.

1 Introduction

Does Machine Translation take context into account? Recent Neural Machine Translation (NMT) models such as Transformers (Maruf et al., 2021) approach the task of sentence-level machine translation without considering the surrounding information beyond the sentence (Bawden et al., 2018). 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) and dialogue translation. Thus, it is still an open research question to what degree these models could learn from the surrounding context.

To answer this question, several context-aware NMT (Voita et al., 2018; Tiedemann and Scherrer, 2017) studies have been conducted by adding surrounding sentences to the models and testing if this helps to capture better specific linguistic phenomena requiring context (e.g. coreference resolution). However, there are only a few researches on discourse or dialogue datasets and context, and many of these studies focus on high-resource languages such as English and German. Therefore, they fail to capture discourse or dialogue phenomena specific to other languages, especially those related to non-Indo-European (IE) and low-resource languages.

For example, one of the specific linguistic phenomena in Japanese discourse is zero pronouns (Nagata and Morishita, 2020) and Honorifics (Feely et al., 2019). Zero pronouns in Japanese are pronouns that are often omitted without indicating any case and gender information, so it is more difficult to find the antecedent. Honorifics in Japanese are exceptional compared to IE languages since they are also a part of verbal morphology having multiple levels of formality for the same meaning (Feely et al., 2019).

In this project, I am focusing on the Japanese language (medium resource) as my primary use case to study two research questions.

- Do current context-aware NMT models (Tiedemann and Scherrer, 2017) with context-enhanced inputs actually help NMT models with understanding the context?
- What kind of context is useful to solve linguistic phenomena for dialogue, such as honorifics?

To answer those questions, this study targets four main contributions.

1. Fine-tuning current state-of-the-art multilingual model mBART (Liu et al., 2020; Tang et al., 2021) with additional context on Japanese Business Scene Dialogue dataset (BSD) (Rikters et al., 2019)

2. Proposing the novel context-aware approach adding three different types of contexts; preceding sentences on source or target side, speaker information, and scene information using BSD dataset

3. Evaluating the context-aware NMT models with CXMI (Fernandes et al., 2021), where the algorithm calculates the information gain between a context-agnostic model and a context-aware model

4. Evaluating the context-aware NMT models with *Hnorifics CXMI* (more detail is described in Section 8.1), which computes CXMI in terms of Honorifics

2 Related Work

2.1 Context-aware MT and Document-level MT

While sentence-level MT translates one sentence in a source-side language into one sentence in a target-side language, document-level MT translates a sequence of sentences in a source-side language into a sequence of sentences in a target-side language in a document. The general approach to the context-aware MT is to combine the preceding sentences (context sentences) on the source side or target side (as discussed in section 2.2), which makes the task similar to document-level MT. However, context-aware MT differs from document-level MT since the goal of context-aware MT is to translate only one sentence which follows the context sentences.

Thus, context-aware MT requires additional challenges compared to existing main MT systems to adapt to additional contexts. It needs a document-aligned parallel dataset for MT that allows the model to combine the preceding sentences sharing the same context, which is a limited amount compared to sentence-level. Evaluation for context-aware MT is also challenging since the current main methods focus on sentence-level evaluation (Rikters et al., 2020). Also, depending on how we combine contexts in the input or the model architecture, adding contexts may cause the models to suffer from understanding long-range dependencies (Yun et al., 2020).

2.2 Context-aware MT Approach

Several methods using Transformers (Vaswani et al., 2017) were proposed for context-aware NMT. The main methods are categorised as 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 also take the preceding sentence(s) as the contexts, however, they use additional encoder modifying the Transformer architecture (Voita et al., 2018; Tu et al., 2018).

Based on the context-aware NMT approaches, some context-aware NMT studies in Japanese were conducted with single-encoder models (Sugiyama and Yoshinaga, 2019; Ri et al., 2021; Rikters et al., 2020). However, there are only a few context-aware NMT with discourse or dialogue datasets in Japanese. Rikters et al. (2020) experimented context-aware MT with source-side factors on Ja-En (Japanese-English) and En-Ja (English-Japanese) discourse datasets. Their approach is to concatenate the preceding sentence(s) from the same document followed by a bos (beginning of sentence) tag <bos> to the source sentence and pass them with a source side factor. The source side factor is the same length as the concatenated tokens, which is additionally provided in training to specify if each token represents context or source sentence, using the symbol of C or S, meaning context or source.

Following those previous studies, this study also uses a single-encoder model since the performance gap between the two models is marginal. It is a relatively simpler architecture without modifying sequence-to-sequence transformer (Sugiyama and Yoshinaga, 2019). To indicate the boundary between the concatenated context-enhanced texts and the input text to be translated, a special separator token </t> is employed in addition to bos and eos (end of sentence) token, without adding the source-side factor.

うなぎが　食べたいな
unagi-ga　tabe-tai-na
eel-OBJ　eat-want-PARTICLE
I feel like eating eel.

Figure 1: Zero pronoun example in Japanese discourse (Ri et al., 2021)

2.3 NMT for Japanese Discourse

In MT in Japanese, specific discourse phenomena such as zero pronouns and honorifics have been one of the main challenges when translated from languages that do not include such phenomena, like English.

2.3.1 Zero Pronouns

Taira et al. (2012) claims that the zero pronouns (omission of subject, object and possessive case) cause decreasing the quality of Ja-En MT. For example, Figure 1 shows a typical zero pronoun, omitting the subject *"I"* completely in the Japanese sentence.

Thus, in Statistical Machine Translation (SMT), zero pronouns were explicitly predicted by incorporating special methods. However, in NMT, the omitted pronouns can be automatically inferred without guaranteeing the correctness of the antecedent (Ri et al., 2021). To handle this issue, several studies are conducted to evaluate context-aware NMT models with hand-crafted contrastive test sets for Japanese to improve the performance in pronoun resolution (Shimazu et al., 2020; Nagata and Morishita, 2020).

2.3.2 Honorifics

Japanese honorifics are special compared with English since 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 (Feely et al., 2019). For example, Feely et al. (2019) explained a typical example of three different types of Japanese honorifics, which is equivalent to *"there are"*. When speaking with family, close friends, or others of equal social status, the informal "ある" (aru) is used. When speaking to superiors, strangers, or older individuals, the polite expression "あります" (arimasu) is used. When expressing deference or humility, the formal expression "ございます" (gozaimasu) is used.

Therefore, formality-aware NMT was proposed by Feely et al. (2019). The experiment manually sets up the formality level to the model beforehand to evaluate honorifics. They evaluate the formality level of the translated sentences by using their formality classifier for both the MT output and the test reference. However, this paper demonstrates honorific assessment without the manual input of the formality level. Instead, it attempts to utilise context-aware NMT, which hypothetically adds context information to select the formality level.

As for datasets, Liu and Kobayashi (2022) constructed KeiCO Corpus, where honorifics are annotated for the Japanese sentences. It contains detailed information on honorific levels, the social relationship between the speaker and the listener, and conversational situations or topics. It aims to improve Japanese honorific-related tasks, including machine translation. However, it cannot be used directly for machine translation since it is not a parallel corpus.

Business Scene Dialogue corpus (BSD) (Rikters et al., 2019) is a parallel corpus in English and Japanese, with honorifics frequently used in Japanese business scenes. Although there is no label for the honorific information, it has speaker and scene information that can indicate the honorific level. The examples are introduced in the paragraph of speaker information in section 5.1. However, to the author's best knowledge, the speaker information and scene information have not been used in the previous studies with the dataset for context-aware NMT approach.

3 Focused Questions

Given the research questions and several problem statements from the related works, this study focuses on the formalised questions below, using metrics of BLEU and COMET to evaluate the overall performance of models and CXMI to evaluate the degree to which they use additional context.

- Does a large language model pretrained on multi-lingual data improve the MT performance (measured by BLEU and COMET scores) for a Japanese discourse dataset?

- To what extent do additional preceding sentences in source and target inputs improve the MT model performance (measured by BLEU and COMET scores)?
- Do additional speaker information and scene information improve the MT model performance (measured by BLEU and COMET scores)?
- Which context-enhanced information (amongst the different numbers of preceding sentences, speaker information or scene information) is actually used to help the model with improving the translation (measured by CXMI)?
- What kind of context-enhanced information is actually used to help the model with improving honorific translation (measured by *Honorifics CXMI*)?

4 Datasets

This experiment uses Business Scene Dialogue corpus (BSD) (Rikters et al., 2019) as a main dataset and AMI Meeting Parallel Corpus (AMI) (Rikters et al., 2020) as a supplemental dataset to compare the performance with the main dataset. The datasets are chosen for the experiment since they are business scene dialogue or meeting recording corpora expected to have complex discourse phenomena. They are document-level parallel corpora consisting of different scenes or meetings and can be translated from both English to Japanese and Japanese to English. Both datasets are publicly available.

In particular, 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). Moreover, each sentence has speaker information that indicates who is speaking. An example of data with the specific structure is explained in the following GitHub link.[1] 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 make a parallel corpus.

[1] https://github.com/tsuruoka-lab/BSD

The BSD data split that this experiment also follows is shown in Table 1. Also, the statistics of the dataset are shown in Table 2.

Table 1: BSD data split

	Training	Development	Test
Sentences	20,000	2051	2120
Scenarios	670	69	69

In Table 2, JA-EN column means the original language is Japanese, and EN-JA column means the original language is English. As shown in the statistics, the data split is balanced with respect to the number of scenes and original languages. The average length of the document is 29.7 sentences, which is relatively short because of the nature of dialogue data compared with text documents such as newspapers.

Table 2: BSD data statistics

	Scene	Docs	Sents	Docs	Sents
		JA-EN		EN-JA	
Train	Face-to-face	122	3525	103	2986
	Phone call	68	1944	75	2175
	General chatting	61	1915	72	1883
	Meeting	56	1964	58	1787
	Training	12	562	19	463
	Presentation	6	607	18	189
	Total	325	10,000	345	10,000
Dev	Face-to-face	11	319	12	314
	Phone call	6	176	7	185
	General chatting	7	223	8	248
	Meeting	7	240	7	219
	Training	1	40	1	23
	Presentation	1	31	1	33
	Total	34	997	35	1054
Test	Face-to-face	12	381	11	345
	Phone call	6	163	7	212
	General chatting	7	221	8	212
	Meeting	7	228	7	229
	Training	1	38	1	30
	Presentation	1	31	1	40
	Total	34	1052	35	1068

Sents is Sentences, Docs is Documents (Scenarios) here.

As for the supplemental dataset AMI, the contents are translations to Japanese from 100 hours of meeting recordings in English. Since the original language of the dataset is English discourse, it contains more short utterances compared with BSD,

such as "Yeah", "Okay", or "Um". The data split this experiment follows is shown in Table 3. The domain and structure of BSD and AMI are similar; however, AMI does not include scene information.

Table 3: AMI data split

	Training	Development	Test
Sentences	20,000	2000	200
Scenarios	30	5	5

5 System Overview

This section explains the model architecture for the context-aware NMT with context-enhanced inputs from English to Japanese. First, how contexts are encoded to generate context-enhanced input representations is discussed. After that, the modification in the model architecture to improve the model's learning from the context-enhanced representations is explained. The implemented code is in the link in the footnote.[2]

5.1 Encoding Context

The primary method of adding context sentences to the model is based on the approach by Tiedemann and Scherrer (2017). There are two types of approaches. The first one is *Extended Source*, where it includes context from the preceding sentences only on the source language to improve the encoder part of the network. The second one is *Extended translation units*, where it increases the segments to be translated, and the larger segments on the source language have to be translated into corresponding units on the target language. For both approaches, only one preceding sentence is added.

Our experiment focuses on the idea of *Extended Source* since we are interested to improve the sentence-level translation performance by separately adding the context on the source side and the target side. In the *Extended Source* approach, to show the boundary between the context from preceding sentences and the current sentence to be translated, the additional sentence-break token *_BREAK_* is inserted as shown in Figure 2.

The following paragraphs provide a more detailed description of the method and motivation

look , Bob ! _BREAK_ - Where are they ?
- Where are they ? _BREAK_ do you see them ?
do you see them ? _BREAK_ - Yes .

Figure 2: Break token in Tiedemann and Scherrer (2017)'s approach

in each different type of context-enhanced information: preceding sentence(s), speaker information and scene information.

Preceding Sentence(s) This study's primary context type is preceding sentence(s) either on the source or target side. The method of adding the context is based on Tiedemann and Scherrer (2017)'s *Extended Source* approach as described above, however we added several modifications. First, the approach is applied not only to the source language but also to the target language too. Second, although Tiedemann and Scherrer (2017)'s approach expanded the sentence only to one preceding sentence, the size of the additional context in our approach *preceding sentence(s)* is chosen from 1 to 4 preceding sentence(s), following to (Fernandes et al., 2021; Castilho et al., 2020). Third, a separator token <*t*> is employed after every context sentence instead of the *_BREAK_* token, which appears only one time as the boundary between the context and the sentence to be translated. The separator token signifies the separation between the context and input sentence and allows the model to learn respective representations. Table 4 and Table 5 show the example inputs with the <t> token.

We compare the context-aware models to the original 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.

Table 4 and Table 5 show the input of context-agnostic model 1-1, source context-aware models (2-1, 3-1, and a part of 4-1), and target context-aware models (1-2, 1-3, and a part of 1-4). When

[2] https://github.com/su0315/discourse_context_mt

the context size is 0, the model is 1-1, and the source sentence is an utterance, *"Have you heard about the new job opening in the sales department?"*, while when the source context size is one so that the model is 2-1, the source sentence is concatenated with the preceding sentence *"I do, what's up?"* with the special token </t>.

As for the experimental setup, the context concatenation is generated separately for each document. Since the dataset consists of separate documents with different business scene conversations, the last sentences from a previous document are irrelevant to the first sentence in the current document. Thus, the first sentence in each document does not concatenate preceding sentence(s) as its additional contexts. For example, in the 4-1 model in Table 4, the first sentence remained to be *"Do you have a moment to talk?"* without concatenating another preceding context on the top of the input sentence since the sentence is the first sentence in the document.

The same system applies to the Japanese target side context-aware model's input in Table 5. 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 the 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 As described in section 4, BSD dataset includes speaker information, so this study tries to utilize them as contexts. Speaker information is useful, especially in dialogue datasets, since there are multiple speakers in a document and each speaker may utter multiple sentences in each turn. Hence, speaker information can inform us when the speaker is changed. The change of speakers can signify a change in discourse style, politeness or even topic distribution. In particular, Japanese honorifics form or causality expression would often be indicated by whom the speaker addresses (Feely et al., 2019).

For the experimental setup, we consider only two speaker types: the one whose input we need to translate, but who may have communicated more sentences in the nearby context (same speaker) and any other speakers who may have spoken within the context window (different speaker) and we do not differentiate from each other. In other words, we only care to encode information about whether there has been a change of speakers within the context. Hence, either a special token <DiffSpeak> (Different speaker) or a <SameSpeak> (Same speaker) is concatenated to each sentence (utterance) of the context in order to differentiate the different speaker in context sentences from the current sentence. Table 6 shows example source inputs with speaker information and with the Japanese translation. The conversation is about a boss and an employee's face-to-face conversation at an office. Here, <SameSpeak> is the boss speaking, and <DiffSpeak> is an employee speaking.

The speaker information would give useful infor-

Table 4: Source context-enhanced inputs with <t> token

Context size (Model Name)	Source sentence(s)
0 (1-1)	Have you heard about the new job opening in the sales department?
1 (2-1)	I do, what's up? </t>Have you heard about the new job opening in the sales department?
2 (3-1)	Do you have a moment to talk? </t>I do, what's up? </t>Have you heard about the new job opening in the sales department?
3 (4-1)	Do you have a moment to talk? ...

Table 5: Target context-enhanced inputs with <t> token (The Japanese sentences below are the reference translation of equivalent English sentences in Table 4.)

Context size (Model Name)	Target sentence(s)
0 (1-1)	販売部の新しい求人の話を聞きましたか？
1 (1-2)	いいですよ、どうしましたか？ </t>販売部の新しい求人の話を聞きましたか？
2 (1-3)	ちょっとお時間良いでしょうか？ </t>いいですよ、どうしましたか？ </t>販売部の新しい求人の話を聞きましたか？
3 (1-4)	ちょっとお時間良いでしょうか？ ...

Table 6: Speaker tag examples with a boss and employee's face-to-face conversation

Context size (model name)	Source side context and current sentence inputs with speaker tags	Translation of current sentence (and preceding contexts) in Japanese
0 (1-1)	<SameSpeak>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 (2-1)	<SameSpeak> Thank you for coming. </t><SameSpeak> First of all, I want to thank you for all your hard work.	(来てくれてありがとう。) 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。
2 (3-1)	<DiffSpeak>Mr Billy, I was told you needed me in your office. </t> <SameSpeak>Thank you for coming. </t> <SameSpeak> 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。 来てくれてありがとう。) 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。
3 (4-1)	<DiffSpeak> Mr Billy, I was told you needed me in your office. ...	(ビリーさん、用があると聞いたのですが。 ...

mation on honorifics in Japanese translation, which English source sentences alone cannot indicate at all. For example, in Table 6, "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。 " (meaning, *"First of all, I want to thank you for all your hard work"*) is a casual ending as in "伝えたい", whereas "ビリーさん、用があると聞いたのですが" (*"Mr Billy, I was told you needed me in your office."*) shows the politeness as in "ですが" at the end of the Japanese sentence. Here, it is not a coincidence that the translation of the boss speaking is a casual expression, and the employee's speaking is a polite expression.

Scene Information As well as speaker information, this study also tries to concatenate scene information called scene tag in BSD dataset, introducing additional special scene tokens. Following BSD dataset scene tags as in Table 4, we prepared six additional tokens; <Face-to-Face>, <Phone call>, <General chatting>, <Meeting>, <Training>, and <Presentation>. One of the tags is added at the very beginning of each source input once to let the model identify what kind of scene discourse it is.

For example, the scene tag of conversation in Table 6 is <face-to-face conversation>, so the 1-1 model's input will be "*<face-to-face conversation> First of all, I want to thank you for all your hard work.*", and 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 hypothetically helps model with understanding the style of the speaking, such as honorifics and casualty, and even scene-specific terminologies frequently used in the scene.

5.2 Context-aware Model Architecture

Baseline All the models for En-Ja translation in this experiment are based on mBART50 (Liu et al., 2020; Tang et al., 2021), with a slight modification for adapting the context-enhanced input representations. mBART is one of the state-of-the-art multilingual NMT models, and the architecture is based on Transformer (Vaswani et al., 2017). It follows BART (Lewis et al., 2020) Seq2Seq pretraining scheme but it is pretrained in 50 languages, including Japanese and English, using multi-lingual denoising auto-encoder strategy to improve its performance. It is pretrained for sentence-level translation (that is, context agnostic 1-1 model in this paper).

Target Context-aware model Architecture To apply the Tiedemann and Scherrer (2017)'s context-aware approach to the target side with training and evaluating the sentence-level translation output, the baseline model architecture needs to be modified since the baseline model receives as the target output the translation of both the context sentences and the current sentence. For example, with a 4-sized context (x-5 model) without further modifications, the model would learn to predict 4 preceding sentences and 1 current sentence on a target side during training. Hence, without any modifications, the loss would be calculated over all 5 sentences

instead of only the current sentence target, which is the goal of sentence-level translation.

To solve the issue, after reading both context sentences and a current sentence, we use the context sentences only as input to the decoder, but we let the target context-aware model pass only the current sentence prediction to the loss function by separating the context sentences and the current sentence. Also, only the current sentence is evaluated during the evaluation. With this regard, the Huggingface[3] implementation on mBART model is slightly adapted as in the code. [2]

Source Context-aware model Architecture The simplest approach is to use directly the baseline model architecture with the extended inputs that include the context encoding as described in the paragraph of Preceding Sentence(s) in Section5.1. However, with the simple approach, it was found that particularly source context-aware models did not outperform the context-agnostic 1-1 model; rather, they decreased the score from the 1-1 model. The column of baseline in Table7 shows the radical score drop in the 2-1 to 5-1 model from the 1-1 model, unlike what was found in (Tiedemann and Scherrer, 2017). This result indicates that the context-aware models with the baseline architecture does not learn effectively from the extended source contexts when tuned on small datasets.

To avoid the problem, the new architecture Source **Co**ntext **Att**ention **Mask** Model (CoAttMask) is proposed for especially source context-aware models to adapt the baseline mBART model for the source context-enhanced inputs. As in Table7, it successfully improved the performance of the baseline model architecture (without CoAttMask) both in no context (1-1) and source context-enhanced inputs (2-1 to 5-1).

The theoretical motivation behind the proposed CoAttMask's architecture is the difference in the decoder's input between the original mBART's pretraining setting and the context-enhanced finetuning setting in this experiment. Baseline model mBART is pretrained to translate one source sentence to one target sentence in much larger datasets compared to BSD dataset, whereas finetuned source context-aware model's goal in this experiment is to translate from several sentences (with context and current sentences) into one target sentence. Without any architectural modification as CoAttMask proposes, the source context-aware models pass 2 to 5 concatenated sentences to the decoder, whereas the original mBART model was pretrained where the decoder was taking only a current sentence. In this case, finetuning mBART with source contexts would become more likely to be finetuning with a slightly different task, and it would need a lot of training samples to adapt to such different tasks. However, BSD dataset is a relatively smaller size, so finetuning with the dataset would not be hypothetically enough for the pretrained mBART model to adapt the context-enhanced source inputs.

In contrast, CoAttMask architecture passes only the current sentence to the decoder instead of passing the current sentence with the concatenated contexts. The mechanism is implemented by masking the attention scores of only source contexts after the context-enhanced concatenated sentences are passed to the encoder before the decoder, as in the yellow block in Figure 3. Since the masking process is implemented after the encoder, the concatenated context sentence information is still represented in the current sentence in the encoder, even though the context sentences themselves are ignored in the decoder by being masked.

Table 7: Performance of CoAttMask model in COMET. **Bold** scores signify the performance improved 1-1 model.

	Baseline	CoAttMask
1-1	0.724	-
2-1	0.661	0.724
3-1	0.665	0.724
4-1	0.662	**0.727**
5-1	0.658	**0.727**

5.3 Hyperparameters

The max input size is set to 128 in all models for padding. In this experiment, truncation is not set so that the model does not cut the current sentence when the context size becomes larger. Training batch size is 4, learning rate is 2e-5, warm up steps is 500, weight decay is 0.01, number of training epochs is 5 for context-aware models and 10 for CXMI random models (which is explained in the

[3] https://huggingface.co/

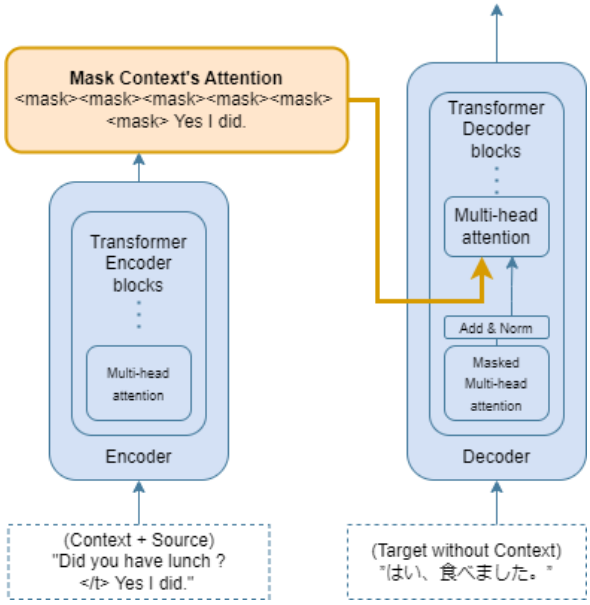


Figure 3: CoAttMask architecture

paragraph on random context size model in Section 6), early stopping patience is 3 for context-aware models and 5 for CXMI random models. All of the parameters and a more detailed setup are in configuration files in the link.[2]

6 Evaluation

6.1 Metrics for Overall Performance

Evaluation of MT outputs is traditionally carried out using the BLEU metric (Papineni et al., 2002), which calculated the n-gram overlap between the MT output and a reference translation. However, BLEU cannot always capture an accurate estimate of the translation taking potentially correct synonyms or paraphrases into account (Smith et al., 2016), since it focuses on counting the number of matching n-grams between the predicted translation and the reference. Thus, BLEU is less efficient in evaluating semantic scores. Especially for context-aware NMT in discourse dataset, BLEU ignores discourse phenomena in the translation of longer pieces of texts (Maruf et al., 2021).

Instead, neural MT evaluation metrics have shown to be more efficient in assessing the quality of MT outputs. COMET (Rei et al., 2020) is a multilingual neural MT evaluation metric proposed to predict MT more accurately within the segment level. Using a multilingual embedding space, it exploits information from both the source-language input and a target-language reference.

This experiment uses BLEU and COMET with the motivation explained above. The training is optimised for the COMET score, and BLEU is calculated only for the purpose of comparing the performance with other studies. While the inputs of BLEU are reference sentences and predicted translations, and it outputs the quality score. the inputs of COMET are source sentences, reference sentences, and predicted translation sentences, and it outputs a quality score ranging from 0 to 1 (Rei et al., 2020). In this experiment, the predicted translation sentence passed to the metrics is only one current sentence without preceding context sentences so that those metrics evaluate one current sentence's translation compared with one equivalent source and one reference sentence.

6.2 Context Evaluation

Although COMET can capture more semantic features than BLEU, it is still difficult to assess whether the context-aware NMTs are improved by the additional context. To indicate it in terms of phenomena-specific aspects, several contrastive datasets are proposed to assess the correct prediction rate in specific linguistic phenomena such as pronoun resolutions. (Bawden et al., 2018; Müller et al., 2018) However, evaluating context-aware NMT on the contrastive dataset is limited in terms of the domain of the text, and it can only assess the model with respect to the limited number of phenomena. The score will depend on the frequency of the phenomena and cannot be generalized to other corpora or inform us on how much and in which cases the model is actually using the additional context (Fernandes et al., 2021). Thus, it would also be hard to develop a future NMT study to reflect the evaluation of context-aware NMT analysis by only depending contrastive dataset.

Therefore, Fernandes et al. (2021) proposed Conditional Cross Mutual Information (CXMI), which assesses how much context-enhanced NMT models actually use the additional contexts in a particular window of context. CXMI measures the entropy (information gain) of a context-agnostic machine translation model and a context-aware machine translation model inspired by the concept of cross-mutual information (XMI) (Bugliarello et al., 2020). This study provides another choice for evaluating context-aware NMT, allowing us to

assess if the context is actually used and the degree of context use on MT with respect to all the context-related phenomena.

Therefore, this study uses CXMI to assess the effectiveness of our context-aware models with context-enhanced inputs. CXMI is calculated between a context-aware model and a context-agnostic model. It is formulated as below by Fernandes et al. (2021). When an additional context is C, target is Y, and source is X, CXMI below measures how much information the context C provides about the target Y given the source X. H_{qMT_A} is the entropy of a context-agnostic machine translation model, and H_{qMT_C} is a context-aware machine translation model. The further detail of the algorithm and implementation method is explained by Fernandes et al. (2021).

$$\text{CXMI}(C \rightarrow Y|X) = H_{qMT_A}(Y|X) - H_{qMT_C}(Y|X, C)$$

When context-aware models improved the score by the additional contexts, CXMI score should be positive. The higher the CXMI is, the more information gain the model exhibits from additional contexts. The illustration of CXMI is in Figure 4.

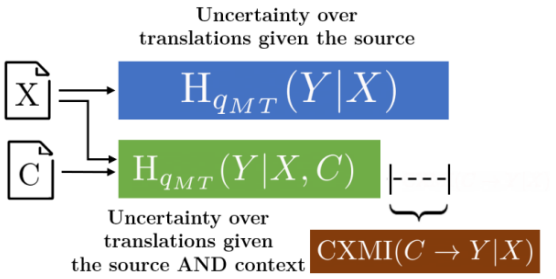


Figure 4: Illustration of how we can measure context usage by a model H_{qMT_A} as the amount of information gained when a model is given the context H_{qMT_C} and sourceH_{qMT_X} vs when the model is only given the H_{qMT_X} (Fernandes et al., 2021)

In this experiment, we compare CXMI scores for context-aware models with preceding sentence(s), speaker information, and scene information. The next paragraph explains the models that need to be prepared for computing CXMI for the experiment.

Random Context Size Model for CXMI CXMI requires a single model that can be tested with both of context-agnostic inputs and context-enhanced inputs since the probability distributions over the prediction on entire vocabularies between the models should be comparable numbers. Thus, context-aware models with a random context size (from 0 to 4) or random context type are required.

This study prepared a 5-1 random context size model for all the source context-aware models and a 1-5 random context size model for all the target context-aware models to calculate CXMI between each context size of a context-aware model and a 1-1 context-agnostic model. The 5-1/1-5 random context size model can predict translation with 0 to 4 source/target context sizes. Additionally, the speaker random model and scene random model for each context size is prepared to calculate CXMI between a speaker information model or scene information model and a model without the speaker or scene information. The random models randomly decide the context size or if the speaker or scene context is combined, in each sentence-level iteration.

7 Experimental Results

7.1 Context-agnostic Models

In Table 8, the result of the baseline 1-1 model in this experiment is compared with previous studies using the same test dataset. Those previous studies used the basic Transformer model with the same max length of 128 (Rikters et al., 2019) with this experiment. It shows that mBART model finetuned in this experiment significantly outperformed the previous studies by more than 12 of BLEU score.

Table 8: 1-1 model score comparison on BSD test data. **Bold** scores signify the best performance.

	BLEU↑	COMET↑
Rikters et al. (2019)	13.53	-
Rikters et al. (2021)	12.93	-
Finetuned mBART	**26.04**	**0.725**

The training data setup for Rikters et al. (2021) is different with their additional dataset.

The result indicates the mBART's unique pretraining scheme may improve En-Ja machine translation. mBART is a multilingual sequence-to-sequence denoising auto-encoder, which is pretreined to denoise the noised input texts (by masking phrases and permuting sentences) across many languages. Also, the pretrained model is a left-to-right autoregressive model that reduces the mis-

match between pretraining and generation tasks, as in bidirectional models such as BERT (Devlin et al., 2019) (Lewis et al., 2020). Thus, the scheme of denoising the corrupted texts on the large multilingual data may give useful cross-language information that is also useful to En-Ja translation, and the pretrained autoregressive model would match the fine-tuned machine translation task (which is also a generation task).

7.2 Context-aware Models

For context-aware models, four types of the models' scores are compared within different context sizes;

- Preceding Sentences Model
- Speaker Information Model
- Scene Information Model
- Speaker & Scene Information Model

To analyze these context-aware models' performance, BLEU and COMET are used for overall performance and CXMI for some models. Each context-aware model is compared with its different baseline model, such as a 1-1 model, "without speaker model", or "without scene model". The details are explained in the following sections.

7.2.1 Preceding Sentences Model

As in Table 9, preceding sentences models are compared with the baseline 1-1 model. The integrated four rows from 2-1 to 5-1 are the source side preceding sentences models, and the following integrated four rows from 1-2 to 1-5 are the target side preceding sentences models. Those context-aware models' scores outperforming the baseline score are highlighted in bold font.

As for the source context-aware models, although 2-1 and 3-1 decreased the baseline scores both in BLEU and COMET, larger context sizes such as 4-1 and 5-1 outperformed the baseline scores. As for target context-aware models, most scores did not outperform the baseline except the BLEU score in the 1-3 model. To explain the performance difference between the source and target context-aware models, it might be caused by the difference in the model architecture between them. For the source context-aware models, CoAttMask model architecture described in Figure 3 is used, whereas the target context-aware models are implemented without it. Thus, there is a possibility that target context-aware model architecture could not adapt the decoder's contex-enhanced inputs as well as CoAttMask model could.

Table 9: Score comparison between preceding sentences models and 1-1 model. **Bold** scores signify the performance improved baseline (BLEU, COMET)

	Model (context size)	BLEU ↑	COMET ↑	CXMI ↑
Baseline	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

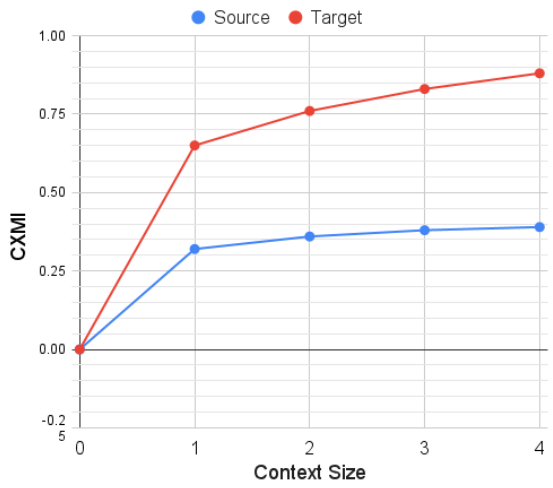
As for CXMI in both source and target context-aware models, all of them are positive scores as in Table 9. The positive scores indicate that the additional contexts are actually used by the models to predict the gold label translation. Furthermore, the scores give a different perspective compared to the scores reported for BLEU and COMET from the result as in Table 9, since the CXMI scores in all target context-aware models are higher even though the target context-aware models' scores in BLEU and CXMI are lower than the baseline model in most of the context size. Also, despite the larger size of source contexts leading the better BLEU, COMET and CXMI scores, the larger size of target contexts causes a slightly lower score in BLEU and COMET and a significantly higher score in a CXMI score, as shown in the 1-5 model in the table.

It might be because the training with target contexts on the small dataset would be cumbersome and cannot converge to the best checkpoint, whereas the target side context is useful to increase the probability distribution on the correct vocabularies. For future work, it is interesting to train the target context-aware models in a larger dataset and compare the performance.

In addition, Figure 5 shows CXMI scores for both the target and source context-aware models in different context sizes. From the figure, it is also common in both of the source side and target side

that the CXMI scores grow up with the context size.

Figure 5: CXMI for source and target context-aware models in each context size



7.2.2 Speaker Information Model

Since the performance of the source context-aware models outperformed the baseline 1-1 model, speaker information is additionally concatenated to them in each context size. As in Table 10, speaker information models with each preceding source sentences are compared with "Without speaker" models. The architecture in all the models is CoAttMask model.

As in Table 11, although "With Speaker" models improved the "Without Speaker" models' BLEU score slightly on 2-1, 3-1, and 5-1, overall speaker information did not improve the score in COMET. Interestingly, except for the shortest context size 2-1 model, CXMI is also negative in all context sizes as in Table 6, showing that the speaker turn information does not encourage the model to use the context information.

From this result, it seems that when used on its own, the speaker turn tags do not improve context usage nor do they provide useful context themselves, however, there is room for improving speaker information in this experiment. The speaker information in this experiment only differentiates if the speaker context sentence is the same or different from the current sentence speaker with two special tokens. However, when there are more than two speakers in the document, it would be useful to enumerate the speaker information as in speaker 1, speaker 2, and speaker 3. Furthermore, the names of speakers, which BSD dataset provides as speaker information, are not utilised in this experiment. Such speaker name information may ideally help the model to correctly translate their pronouns or the speaker (listener)'s names in the target side language, which can be challenging. For example, in the third row from the bottom in Table 15, when the source sentence is *"So, it's Mr. Tada?"*, both 1-1 and 3-1 model translate the wrong family name such as 戸倉（Tokura） and 戸田 (Toda), even though the correct name in the reference is 多田 (Tada).

Table 10: Score comparison between with and without speaker information. **Bold** scores signify the best performance (BLEU, COMET) for each context size.

Model (context size)	Without Speaker		**With Speaker**		CXMI ↑
	BLEU ↑	COMET ↑	BLEU ↑	COMET ↑	
2-1 (1)	25.87	0.724	**25.94**	0.718	0.04
3-1 (2)	25.41	0.724	**26.09**	0.72	-0.01
4-1 (3)	26.09	0.727	26.03	0.722	-0.02
5-1 (4)	26.09	0.727	**26.39**	0.726	-0.01

7.2.3 Scene Information Model

As well as speaker information model, scene information models are also compared with "Without scene" models in each different source context sizes in Table 11. The architecture is also the same as speaker information models. However, unlike the speaker information model, scene information model can be added when the context size is zero too, since it does not need preceding sentences unlikely in speaker information, where it indicates if the speaker of preceding sentences is the same or different as the current sentence.

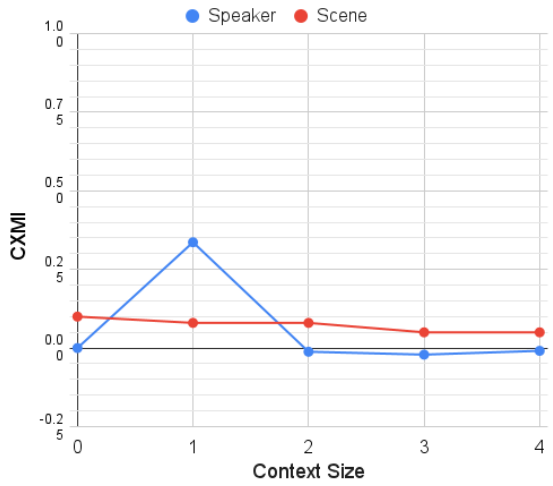
In contrast to speaker information models, "With Scene" models outperformed "Without Scene" models in BLEU and COMET on all of the context sizes. Besides, CXMI is positive in all context sizes with a decrease when the context size is larger. Figure 6 also shows the comparison in CXMI score for speaker information models and scene models. Scene information models show stable CXMI scores in all context sizes, whereas speaker information shows a drastic decrease after context size 1. This result indicates additional scene information, even with preceding source sentences, improves

the model's performance, however, the extent of scene information use in terms of predicting correct translation is slightly decreased as the context size becomes larger.

Table 11: Score comparison between with and without scene information. **Bold** scores signify the best performance (BLEU, COMET) for each context size. Underlined scores signify the best value for each metric.

Model (context size)	Without Scene		With Scene		CXMI ↑
	BLEU ↑	COMET ↑	BLEU ↑	COMET ↑	
1-1 (0)	26.04	0.725	**26.19**	**0.726**	0.10
2-1 (1)	25.87	0.724	**26.18**	**0.727**	0.08
3-1 (2)	25.41	0.724	**26.26**	**0.727**	0.08
4-1 (3)	26.09	0.727	**26.27**	**0.731**	0.05
5-1 (4)	26.09	0.727	**26.10**	**0.728**	0.05

Figure 6: CXMI for speaker and scene model in each context size



7.2.4 Speaker and Scene Model

Given all the results so far and Table 11, the 4-1 model (source context size 3) with scene information shows the best COMET score of 0.731. Therefore, to investigate if adding both scene and speaker information would even improve the scene information model, BLEU and COMET scores of scene & speaker models were also compared with scene models in Table 12. The result shows that scene & speaker models in 2-1 and 3-1 outperformed scene models in COMET score. Also, the 3-1 scene & speaker model performed the best COMET score 0.740 amongst all of the context-aware models, outperforming the 4-1 scene model of 0.731.

Interestingly, while speaker information on its own did not outperform the model without speaker information, the combination of speaker information and scene information outperformed the model without them. It indicates that certain types of scenes have related information in some way to speaker information, which makes the speaker information useful by being combined with scene information.

As described in Table 2, there is some scene information where speaker information may become more useful. For example, most of the speakers in the presentation scenes may be the one presenting without switching the speaker often, whereas several speakers may be expected to switch frequently in meeting scenes. In that case, there is a possibility that the speaker and scene information model can converge more effectively with the regularised frequency of the speaker switch by the scene information. To further research, it would be interesting to analyse the relationship between the speaker switch frequency and the scene type in future work.

Table 12: Score comparison between with scene and with scene & speaker. **Bold** scores signify the best performance (BLEU, COMET) for each context size. Underlined score signifies the best value for each metric.

Model (context size)	Scene		**Scene & Speaker**	
	BLEU ↑	COMET ↑	BLEU ↑	COMET ↑
2-1 (1)	26.18	0.727	26.18	**0.730**
3-1 (2)	26.26	0.727	**26.41**	**0.740**
4-1 (3)	26.27	0.730	26.07	0.730
5-1 (4)	26.10	0.728	**26.15**	0.720

Best Contexts for BSD Dataset As in Table 13, compared with the completely context-agnostic baseline score (1-1), which is 0.725 in COMET, the best context-aware model with preceding 2 sentences, scene, and speaker information shows a significant improvement by 0.15. Also, for comparison with other studies, it is noted with the improvement of BLEU score to 26.41 from 26.04. However, training in this experiment is tuned to optimise COMET, so if the experiment is optimised for BLEU, greater improvement in BLEU can be expected.

Table 13: Score comparison with the best context-aware models and context-agnostic model. **Bold** scores signify the best performance (BLEU, COMET).

	BLEU ↑	COMET ↑
Context-agnostic (1-1)	26.04	0.725
Preceding 4 sentences (5-1)	26.09	0.727
Preceding 3 sentences (4-1) + Scene	26.27	0.731
Preceding 2 sentences (3-1) + Scene + Speaker	**26.41**	**0.74**

7.3 Performance in a Similar Dataset

From the result above, it is found that a larger size of source context, such as 5-1, 4-1 with scene information, and 3-1 with scene and speaker information are the best performances with BSD dataset, however, it is questionable if the trend applies to other datasets which have a similar domain. To answer this question, this study also attempted to add the contexts using AMI dataset, which is also En-Ja dataset with a similar domain as introduced in Section 4. For the additional contexts for AMI dataset, only the preceding sentences in the smallest context size and largest context size are considered, since AMI dataset does not provide scene information as explained in Section 4.

Table 14: Score comparison between 1-1 model and context-aware models with AMI dataset and BSD dataset. (AMI test data is used for the model for AMI, and BSD test data is used for the model for BSD. **Bold** scores signify the best performance (BLEU, COMET) for each context size.)

Model		AMI		BSD	
		BLEU ↑	COMET ↑	BLEU ↑↑	COMET ↑
Baseline	1-1	32.46	0.852	26.04	0.725
Source context	2-1	**32.80**	**0.858**	25.87	0.724
	5-1	32.05	0.846	**26.09**	**0.727**
Target context	1-2	32.13	0.848	25.85	0.720
	1-5	**32.56**	0.850	24.96	0.710

Table 14 shows the score comparison between AMI dataset and BSD dataset in a certain context size of the models. First of all, the performance of AMI is entirely better than BSD. Even in 1-1, the AMI scores are much higher than BSD scores. Secondly, the type of context-aware model outperforms the 1-1 model is different in AMI dataset. While the model trained with BSD dataset outperforms the baseline with the context of preceding four source sentences (5-1 model), AMI does not. Instead, the model trained with AMI dataset outperforms the baseline with the context of the preceding one source sentence (2-1) and preceding four target sentences (1-5).

From this result, it is found that different datasets require appropriate context size and context location to improve performance, even when the language, domain, and size of the dataset are the same or similar. An observation to explain the observed difference is that the obtained baseline scores on AMI dataset are significantly higher, hence it is possible that for data where the context-agnostic MT model has high performance, there is little to be gained by the additional context. However, more experiments and analysis would be necessary to confirm this hypothesis.

8 *Honorifics CXMI* and Linguistic Analysis

8.1 *Honorifics CXMI*

To evaluate how much additional context is actually used to improve honorifics, this study also attempted to compute CXMI for token-level honorific expressions, which this paper refers to as *Honorifics CXMI*. *Honorifics CXMI* calculates CXMI on the probability distribution only where its gold label is an honorific expression. Whereas the normal CXMI is calculated per sentence (summing up each token-level score within the sentence) and averaged over the number of sentences, *Honorifics CXMI* score is calculated for each token and averaged over the number of the honorific tokens in the full dataset. Thus, CXMI and *Honorifics CXMI* are not comparable in the score of each other.

Inspired by Japanese honorific word lists proposed by Fernandes et al. (2021) and Farajian et al. (2020), the following tokens are selected as the main honorific expressions with modification to collaborate with the mBART50tokenizer (Liu et al., 2020; Tang et al., 2021) that this experiment uses; "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

Table 15: Comparison between a context-agnostic model (1-1) and a context-aware model (3-1) in predicting honorific token "伺”. (Underlined words signify that the 3-1 model improved the 1-1 model in predicting the correct token.)

Speaker	Source Sentence	Reference Sentence	1-1 Model Prediction	3-1 Model Prediction
1	Then can I drop by tomorrow afternoon?	それでは、明日の午後にでも伺いたいのですが。	では、明日の午後に寄ってもいいですか?	では、明日の午後に寄ってもいいですか?
1	When is a good time?	お時間どうしましょう？	いつがいいですか?	何時がいいですか?
2	How about getting you to come around 5 o'clock in the afternoon?	じゃあ、午後５時くらいに来てもらえますか？	では、午後5時ごろに集まったらどうでしょうか?	では、午後5時ごろに集まったらどうでしょうか?
1	Okay.	了解致しました。	分かりました。	わかりました。
1	Sorry, who should I look for?	恐縮ですが、どなた様宛に伺えばよろしいでしょうか？	すみません、誰を探したらいいですか?	すみません、どなたがいいですか?
2	Oh, you can just look for me.	ああ、私宛でいいですよ。	あ、私のところ探してみて下さい。	あ、私探していただけますよ。
2	My name is Tada and I'm the assistant inspector.	警部補の多田と言います。	私、戸田と申します、検察のアシスタントです。	私、戸田と申します、アシスタント検収員です。
1	So it's Mr. Tada?	多田様でいらっしゃいますね？	では、戸倉さんですか?	では、戸田さんですか?
1	Okay.	承りました。	分かりました。	わかりました。
1	I, Takada from Company I will go to your place at 5 o'clock in the afternoon tomorrow.	明日の午後５時に、わたくし、I社の高田が伺います。	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時に御社へお伺いします。

categorized either three types of honorifics; respectful (sonkeigo, 尊敬語), humble (kenjogo, 謙譲語), polite (teineigo, 丁寧語), and the tokens above are merely a part of them. There are much more types of honorifics expressions, such as word beautification (bikago, 美化語), and courteous language (teichogo, 丁重語) (Liu and Kobayashi, 2022), however, this study only used the listed honorifics expressions above, since many of other tokens can be non-deterministic that can either be honorifics or non-honorifics expression depending on the surrounding contexts and situations so that the tokenizer cannot tokenize them correctly. This study merely attempts to examine if *Honorifics CXMI* gives us useful information to analyze how much contexts are used to predict correct honorifics in a computational approach.

8.2 *Honorifics CXMI* and Linguistic Analysis

Table 16 shows the result of *Honorifics CXMI* and maximum scores of the *Honorifics CXMI* in the test data in each context-aware model. The first column describes the type of context-aware models, the second column shows the *Honorifics CXMI* scores between the context-aware models and the context-agnostic model 1-1, and the third column shows the maximum *Honorifics CXMI* score with the gold label honorific token that should be ideally predicted in the max score location. The honorifics token shows the honorific expression in the specific location, where the largest information amongst all honorific tokens in the dataset from the additional contexts is used to predict the correct honorific token .

Table 16: *Honorifics CXMI* between source side preceding sentences models and 1-1 context-agnostic Model

	Honorifics CXMI ↑	Max *Honorifics CXMI* Score ↑ (Token)
2-1	0.05	1.53 (伺/ukaga)
3-1	0.07	2.05 (伺/ukaga)
4-1	0.06	2.19 (伺/ukaga)
5-1	0.06	2.47 (伺/ukaga)

According to the *Honorifics CXMI* in the first column, honorific scores do not relate to the context size, however, all context-aware models with the preceding sentences show positive scores. It indicates that the model actually used a certain amount of information to predict the correct honorific translation from the additional contexts.

Interestingly, as in the max *Honorifics CXMI* scores in the second column, all the context-aware models from 2-1 to 5-1 show the identical max score token "伺 (ukaga)" in a specific location in the dataset. It means the prediction of "伺 (ukaga)" benefits more from the context in context-aware models with preceding sentences. The honorific token "伺 (ukaga)" is a token that is a component of "伺う(ukagau)", which is a verb meaning "go" or "ask" in Japanese honorific expression. In particular, "伺う(ukagau)" is categorized as a humble (kenjogo, 謙譲語) expression, and it is one of the most formal expressions used in a business email or very formal speech (Liu and Kobayashi, 2022).

Since humble (kenjogo, 謙譲語) is used when the speaker addresses him/herself to respect the hearer by lowering the position of oneself (Rahayu, 2013), "伺う(ukagau)" should be used strictly by the speaker to address him/herself.

Table 15 shows the conversation from the test data that includes the token "伺 (ukaga)" which obtained the highest *Honorifics CXMI* score among the rest of the tokens examined (highlighted by the underline in the bottom row). The scene of the conversation is phone call, and the topic is making appointments for requests to be present at the general meeting of shareholders. Each column compares the source sentence, gold label reference, the context-agnostic 1-1 model, and the 3-1 model's prediction[4], which is one of the models having high *Honorifics CXMI* score (2.05) for the "伺 (ukaga)" token (as shown in Table 16).

Note that as shown in the dialogue of Table 15, even with the context-aware models that do improve the translation with respect to honorifics, still a lot of instances are mistranslated with respect to honorifics. In the first row (with speaker 1) with the English source sentence of *"Then can I drop by tomorrow afternoon?"*, the Japanese reference sentence includes the honorific expression "伺 (ukaga)" which is equivalent to "drop by" with a connotation of being humble; nevertheless, neither the 1-1 model nor 3-1 model predictions include the "伺 (ukaga)" token; instead, they predict "寄って", which is semantically equivalent to "drop by", however, it is not a honorific expression showing humbleness (kenjogo, 謙譲語). Similarly, in the fifth row with speaker 1 and the English source sentence of *"Sorry, who should I look for?"*, the Japanese reference sentence has the honorific expression "伺 (ukaga)" in the second line, which is equivalent to "look for" in the English source sentence; however, both 1-1 model and 3-1 model did not predict the token correctly, instead 1-1 model predicts "誰を探したら探したらいいですか?" (meaning "Who should I look for" without honorific expression), and 3-1 model predicts "どなたがいいですか?" (meaning "Which person is better?", which is not semantically precise).

In contrast, in the very bottom row with speaker 1 and the English source sentence of *"I, Takada from Company I will go to your place at 5 o'clock in the afternoon tomorrow"*, the context-aware model shows improvement from the context-agnostic model. While the 1-1 model does not correctly predict "伺 (ukaga)", as in "申します" (meaning, "I'm (Takada)", which is semantically not appropriate), the 3-1 model finally and correctly predicts the token within the verb "お伺いします" (Oukagaishimasu, *"(I) go to your place"*).

To examine if additional speaker and scene information is actually used to predict the correct honorific, the *Honorifics CXMI* between the models with preceding sentences + speaker & scene information and the model with only preceding sentences were also calculated as in Table 17. However, according to the result, most of the speaker & scene model shows negative scores in *Honorifics CXMI*, meaning additional speaker and scene information to each context size of preceding sentences did not help the model with predicting the correct honorifics.

Since the speaker and scene information shows the best overall performance as described in Table 13 and the CXMI scores are positive, the negative scores of *Honorifics CXMI* show a different perspective on the result. It indicates the additional information of speaker and scene might be useful for other context but not to predict correct honorifics in BSD dataset. This could be attributed to the fact that BSD dataset always includes honorifics regardless of the type of the scenes hence the scene token may not be informative. Also, the speaker information only signifies the information of the speaker switch in each utterance, which might not be enough to indicate the relationship between the speaker and the listener, which would be useful information to predict the correct honorific. Since we found that there is still room for improvement in honorifics translation, further work trying to better encode the speaker relation information could help improve the current performance.

9 Discussion

From this study, several findings are discovered. In particular, scene information and the larger size of source side contexts improved the English-Japanese translation performance in the experiment.

[4]For calculating CXMI, a single random context size model is used for both the context-aware model and context-agnostic model. The translated outputs in Table15 are predicted by 1-1 model and 3-1 model, which is specifically trained to predict with the fixed context size.

Table 17: *Honorifics CXMI* between preceding sentences + speaker & scene models and only preceding sentences models (2-1, 3-1, 4-1, and 5-1)

	Honorifics CXMI ↑
2-1 speaker & scene	-0.04
3-1 speaker & scene	0.00008
4-1 speaker & scene	-0.007
5-1 speaker & scene	-0.01

Also, according to their positive CXMI scores, additional context-enhanced information is actually used in those best-performance models. For the Japanese discourse phenomena such as honorifics, source context-enhanced information helped the model with translating the honorifics according to the *Honorifics CXMI* score.

However, the best type and size of context is merely for the specific data used in this experiment, and it still cannot be generalized for other experimental setups and datasets. In general, context-aware NMT studies tend to propose only one or two preceding sentences for source side context both in high-resource language and Japanese language (Tiedemann and Scherrer, 2017; Voita et al., 2018; Rikters et al., 2020; Ri et al., 2021; Nagata and Morishita, 2020) and some of them fail to show the significant improvement in context-aware models. Given the previous studies, this study proposes the potential improvement by larger context sizes, such as 4-1 and 5-1, even when shorter context sizes, such as 2-1 and 3-1, did not outperform the context-agnostic model. As for CXMI, the study by Fernandes et al. (2021) indicated that increasing the context size from two sentences decreased the CXMI score in English to German Translation, whereas our experiment result gave the opposite result. It also shows the relation between the number of preceding sentences and the degree of the use of context for correct translation may differ in language, dataset, domain, or even other factors that cannot be explained by this study.

10 Conclusion and Future Work

This paper attempted to analyse if the additional context-enhanced inputs improve the current NMT models performance in English-Japanese dialogue translation, and what kind of contexts give useful information. For the additional context-enhanced inputs, different types of contexts (preceding sentences, speaker information, scene information, and both speaker and scene information) are concatenated with the original input, which is the novel approach for context-aware NMT with BSD dataset. To effectively utilize the source context, CoAttMask model architectures are proposed to adapt the context-enhanced inputs to mBART pre-training scheme without contexts. For evaluating how much context is actually used to predict correct translation and correct honorific translations, this experiment computed CXMI and *Honorifics CXMI*.

With the proposed context-enhanced inputs and the model architecture, we were able to tune mBART on a small dataset in a context-aware approach and obtain improved MT performance. Especially source side preceding sentences models with larger context size and scene information showed the highest scores, and the best model with preceding 2 sentences, scene information and speaker information outperformed the baseline by 0.37 in BLEU and 0.015 in COMET. Also, most of the context-aware models except speaker information show positive scores in CXMI and *Honorifics CXMI*, meaning those models actually used the additional contexts to predict correct tokens. Furthermore, *Honorifics CXMI* gave us useful information to analyse prediction on honorific expression. From the results, source side preceding sentences help the model with predicting the correct honorific expressions, and the token which obtains the maximum *Honorifics CXMI* indicates a certain honorific token that benefits more from the context-aware model than other honorific expressions. However, it is also noted that increasing CXMI scores does not always improve the overall performance of MT measured by BLEU and COMET.

For future work, several related studies would be explored. First, exploring the relationship between speaker information and scene information in terms of the number of speakers and the frequency of the speaker switch would be helpful in improving the representation of speaker information. Second, analyzing other challenging discourse phenomena in Japanese (Nagata and Morishita, 2020; Feely et al., 2019; Taira et al., 2012; Ri et al., 2021; Liu and Kobayashi, 2022; Shimazu et al., 2020; Nagata and Morishita, 2020) with CXMI (Fernandes et al., 2021), or even in other languages with the

context-aware approach (Voita et al., 2018; Tiedemann and Scherrer, 2017; Fernandes et al., 2021; Castilho et al., 2020; Yin et al., 2021; Miculicich et al., 2018), would give a relatively more generalised interpretation of what kind of context is useful. Third, training the model in several larger datasets with a similar domain or different domains would allow context-aware models to gain more information than those trained with BSD dataset (Rikters et al., 2019). It would also be interesting to attempt transfer learning, training the model in larger datasets such as ISWLT (Cettolo et al., 2017) and WMT (Barrault et al., 2020), finetuning it with BSD dataset, and then comparing the test result in BSD dataset in this study. Especially it would be interesting to examine if the target context-aware model's performance, which did not outperform the baseline model, would show a similar or different trend in such large datasets. Lastly, experimenting with prompt learning would also be interesting as it would allow to study the impact of context-enhancing on recent state-of-the-art large language models (LLMs) (Peng et al., 2023; Taori et al., 2023; Touvron et al., 2023). Along these lines we could adapt our problem setup to examine whether context-enhanced prompts would affect the quality of LLMs outputs.

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 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.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 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.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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 the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

Weston Feely, Eva Hasler, and Adrià de Gispert. 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.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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.

Muxuan Liu and Ichiro Kobayashi. 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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Masaaki Nagata and Makoto Morishita. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 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.

Ryokan Ri, Toshiaki Nakazawa, and Yoshimasa Tsuruoka. 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.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 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.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 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.

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

Sho Shimazu, Sho Takase, Toshiaki Nakazawa, and Naoaki Okazaki. 2020. Evaluation dataset for zero pronoun in Japanese to English translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3630–3634, Marseille, France. European Language Resources Association.

Aaron Smith, Christian Hardmeier, and Joerg Tiedemann. 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.

Amane Sugiyama and Naoki Yoshinaga. 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.

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero pronoun resolution can improve the quality of J-E translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 111–118, Jeju, Republic of Korea. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Jörg Tiedemann and Yves Scherrer. 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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 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.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? 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 788–801, Online. Association for Computational Linguistics.

Hyeongu Yun, Yong keun Hwang, and Kyomin Jung. 2020. Improving context-aware neural machine translation using self-attentive sentence embedding. In *AAAI Conference on Artificial Intelligence*.