General Paper

Japanese–English Conversation Parallel Corpus for Promoting Context-aware Machine Translation Research

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

Most machine translation (MT) research has focused on sentences as translation units (sentence-level MT), and has achieved acceptable translation quality for sentences where cross-sentential context is not required in mainly high-resourced languages. Recently, many researchers have worked on MT models that can consider a cross-sentential context. These models are often called context-aware MT or document-level MT models. Document-level MT is difficult to 1) train with a small amount of document-level data; and 2) evaluate, as the main methods and datasets focus on sentence-level evaluation. To address the first issue, we present a Japanese–English conversation corpus in which the cross-sentential context is available. As for the second issue, we manually identify the main areas where sentence-level MT fails to produce adequate translations in the lack of context. We then create an evaluation set in which these phenomena are annotated to alleviate the automatic evaluation of document-level systems. We train MT models using our corpus to demonstrate how the use of context leads to improvements.

Key Words: Parallel Corpora, Machine Translation, Manual Annotation

1 Introduction

The quality of machine translation (MT) for written text and monologues has vastly improved because of the increased amount of available parallel corpora that are required to train MT models and recent advances in neural network technologies that enable efficient use of the available corpora. However, there is much scope for improvement in the context of dialogue or conversation translation. One typical case is the translation from a pro-drop language to a non-pro-drop language, where correct pronouns must be supplemented according to the context. The omission of pronouns occurs more frequently in spoken language than in written language. Recently, to solve this type of problem context-aware MT models have attracted considerable research attention (Tiedemann and Scherrer 2017; Voita et al. 2019).

However, there are almost no parallel conversation corpora with context information, except for the noisy Open Subtitles corpus (Tiedemann 2016). Most of the available parallel corpora are

† The University of Tokyo
shuffled, and sentences do not remain in the original order as in the source material. Some of the corpora are automatically filtered to improve quality, which means that even if they were in the original order, there would be gaps between consecutive sentences. Lastly, not all unfiltered and unshuffled corpora have document boundaries. All these factors hinder the training of context-aware models.

Thus, a parallel document- and sentence-aligned conversation corpus would be advantageous in taking MT research in this field to the next stage. In this paper, we introduce a newly constructed Japanese–English conversation corpus that contains three sub-corpora: business scene dialogue (BSD) (Rikters et al. 2019), Japanese translation of AMI meeting corpus (AMI\(^1\)) (McCowan et al. 2005), and the Japanese translation of OntoNotes 5.0 (ON\(^2\)) (Weischedel et al. 2011). The corpus contains multiperson conversations in various situations: business scenes, meetings under specific themes, broadcast conversations, and telephone conversations. Our main contribution is the production and release of these datasets, their analysis, and markup of errors caused by translating context-dependent sentences using context-agnostic MT systems.

The structure of this paper is as follows: We summarise related work in Section 2 and in Section 3 we explain how the corpus was constructed. In Section 4, we investigate the shortcomings of modern MT with regard to translating conversations. Section 5 provides an overview of our MT experiments, and in Section 6 we present our human evaluation approach and results. Finally, we conclude the paper in Section 7 and introduce plans for future work.

2 Related Work

Modern MT systems are based on neural networks trained with parallel sentences between languages, that is, parallel corpora. Most of the existing ready-to-use parallel corpora comprise written text obtained from Web crawl, patents (Goto et al. 2011), and scientific papers (Nakazawa et al. 2016). Even though some parallel corpora do comprise spoken language, they are generally monologues (Cettolo et al. 2012; Di Gangi et al. 2019) or contain a lot of noise (Tiedemann 2016; Pryzant et al. 2018). Most of the MT evaluation campaigns such as WMT\(^3\) and WAT\(^4\) adopt the written language, monologue, or noisy dialogue parallel corpora for their translation tasks. Among these corpora, there is only one clean, dialogue parallel corpus (Salesky et al. 2018)

\(^{1}\) AMI corpus - \url{http://groups.inf.ed.ac.uk/ami/corpus/index.shtml}  
\(^{2}\) OntoNotes Release 5.0 - \url{https://catalog.ldc.upenn.edu/LDC2013T19}  
\(^{3}\) \url{http://www.statmt.org/wmt20/}  
\(^{4}\) \url{http://lotus.kuee.kyoto-u.ac.jp/WAT/}
adopted by IWSLT\(^5\) for the conversational speech translation task.

JParaCrawl (Morishita et al. 2020) is a recently announced large English–Japanese parallel corpus built by crawling the web and aligning parallel sentences. Its size is impressive, but it is composed of noisy web-crawled data and contains many duplicate sentences. Unlike our corpus, JParaCrawl does not have meta-information or access to cross-sentential context.

Translation that requires contextual information has received much attention in recent years. Voita et al. (2019) evaluated the limitations of modern MT systems when translating English to Russian; they constructed new development and evaluation sets based on human evaluation. The sets target linguistic phenomena: deixis, ellipsis, and lexical cohesion. Our dataset does not aim to evaluate specific discourse phenomena; however, the development/evaluation sets contain complete documents of consecutive sentences, which allow the analysis of discourse phenomena involved in translation.

3 Corpus Description

Our corpus consists of three sub-corpora, each of which originates from different sources: BSD, AMI, and ON. The number of sentence pairs from each corpus is shown in Table 1. We regard sentence pairs as unique if at least one of the source and target sentences are different. For example, if we have two identical sources, but they have different translations, the pairs are considered unique. We provide balanced development and evaluation splits from only the BSD sub-corpus, as it is the least noisy part in our corpus.

BSD has been newly constructed, whereas AMI and ON are translations of the existing English versions of these corpora. Even though BSD has the highest quality of the three, we use all three,

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Development</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSD</td>
<td>JA→EN</td>
<td>1,257</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>EN→JA</td>
<td>1,248</td>
<td>35</td>
</tr>
<tr>
<td>AMI</td>
<td>JA→EN</td>
<td>171</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>EN→JA</td>
<td>46</td>
<td>14,075</td>
</tr>
<tr>
<td>ON</td>
<td>Broad</td>
<td>14,354</td>
<td>1,052</td>
</tr>
<tr>
<td></td>
<td>Tele</td>
<td>11,117</td>
<td>1,068</td>
</tr>
</tbody>
</table>

Table 1 Statistics of the constructed conversation parallel corpus. JA→EN represents scenarios that are written in Japanese then translated into English and EN→JA represents scenarios constructed in the reverse way.

\(^5\) http://workshop2019.iwslt.org
as it is beneficial to have more data for training. Because AMI and ON have similar domains, they allow a meaningful increase in the total relevant training dataset size for experiments. We also conducted several ablation experiments (end of Section 5.2), showing that removing AMI and ON from the training data decreases MT quality. In the following subsections, we describe the construction process of each sub-corpus and its contents.

3.1 Business Scene Dialogue

The BSD sub-corpus was constructed without the use of any pre-existing resources. We choose business conversation as the domain of the corpus because the business domain is neither too specific nor too general, and we believe that a clean conversational parallel corpus is useful for opening new directions for MT research. We hope that this corpus will become a standard benchmark dataset for machine translation.

The three steps in the construction process are as follows: 1) selecting business scenes, 2) writing monolingual conversation scenarios according to the selected scenes, and 3) translating the scenarios into the other language, each of which is explained in detail in the following subsections.

There were a total of 23 English scenario writers, 28 Japanese scenario writers, 13 EN→JA translators, and 8 JA→EN translators, along with three overall translation checkers. Some translators were also scenario writers.

This corpus is unique in that each scenario is annotated with scene information, as shown at the top of Table 2. In conversations, utterances are often very short and vague; therefore, it is possible that utterances will be translated differently depending on the situations in which the conversations take place. For example, the Japanese expression “すみません” can be translated into several English expressions such as “Excuse me.” (when you call a store attendant); “Thank you.” (when you receive gifts); or “I’m sorry.” (when you need to apologise). Using the scene information, it is possible to differentiate the translations, which is difficult to do with only contextual sentences. Furthermore, it might be possible to connect scene information to the multimodal translation, such as estimating scenes using visual information, which has rarely been studied.

1) Business Scene Selection

The business scenes were carefully selected to cover a variety of business situations, including meetings and negotiations, as well as so-called “water-cooler” chats. The details are presented in Table 3. Although most of the scenes are clearly distinct, “face-to-face” and “general chatting” may seem very similar at first. “Face-to-face” represents work-related conversations in the
Table 2  An example of the Japanese–English business conversation parallel corpus. Scene: phone call; title: customer inquiry.

<table>
<thead>
<tr>
<th>Scene</th>
<th>Scenarios JA→EN</th>
<th>Sentences JA→EN</th>
<th>Scenarios EN→JA</th>
<th>Sentences EN→JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>face-to-face</td>
<td>535</td>
<td>16,481</td>
<td>458</td>
<td>14,858</td>
</tr>
<tr>
<td>phone call</td>
<td>279</td>
<td>8,720</td>
<td>256</td>
<td>7,770</td>
</tr>
<tr>
<td>general chatting</td>
<td>233</td>
<td>7,674</td>
<td>239</td>
<td>7,372</td>
</tr>
<tr>
<td>meeting</td>
<td>224</td>
<td>7,647</td>
<td>265</td>
<td>8,952</td>
</tr>
<tr>
<td>training</td>
<td>37</td>
<td>1,379</td>
<td>47</td>
<td>1,549</td>
</tr>
<tr>
<td>presentation</td>
<td>17</td>
<td>499</td>
<td>53</td>
<td>1,899</td>
</tr>
<tr>
<td>sum</td>
<td>1,325</td>
<td>42,400</td>
<td>1,318</td>
<td>42,400</td>
</tr>
</tbody>
</table>

Table 3  Detailed statistics of the BSD sub-corpus.

business scene where people talk directly to each other. Conversely, “general chatting” refers to more of a “water-cooler” conversation between colleagues, which is far less formal and not always business-centred. The “meeting” scene represents situations where the conversation is more one-to-all or all-to-all as opposed to “face-to-face,” which would generally be one-to-one. We include “general chatting” in business scenes because it is useful in situations such as business-related parties.

We also took care not to select specialised scenes that are suitable only for a limited number of industries. We ensured that all selected scenes were generic to a broad range of industries.
2) Monolingual Dialogue Scenario Writing

Dialogue scenarios were written monolingually for each of the selected business scenes. Half the monolingual scenarios were written in Japanese and the other half were written in English (42,400 sentences for each language). This is because we aim to cover a wide range of lexicons and expressions for both languages in the corpus. Writing the scenarios in only one language may not incorporate useful, important expressions in the other language when they are translated in the following step.

3) Scenario Translation

The monolingual scenarios were translated into the other language by human translators. They were asked to make the translations not only accurate but also as fluent and natural as possible. This principle is adopted to eliminate several common tendencies of human translators when performing Japanese–English translation on a written text. For example, Japanese pronouns are usually omitted in a dialogue; however, when English sentences are literally translated into Japanese, translators tend to include unnecessary pronouns. It is acceptable as a written text but rather unusual as a spoken text.

Quality Control

The scenario writing and translation were performed by several scenario writers and several translators; therefore, quality control became necessary. In addition, it is important to check whether the scenarios are suitable for a business scene. To guarantee quality and suitability, we asked one person to supervise the entire construction process and regulate its execution. The person satisfies the following conditions to guarantee that the conversations are natural:

- has the experience of being engaged in language learning programs, especially for business conversations
- is able to smoothly communicate with others in various business scenes in both Japanese and English
- has the experience of being involved in business

3.2 AMI Meeting Parallel Corpus

The original AMI Meeting Corpus is a multimodal dataset containing 100 hours of meeting recordings in English. The parallel version was constructed by asking professional translators to translate the utterances from the original corpus into Japanese. As the original corpus consists of speech transcripts, the English sentences contain many short utterances (e.g., “Yeah” and
“Okay”) or fillers (e.g., “Um”), and these are translated into Japanese as well. Therefore, the AMI sub-corpus contains many duplicates at the sentence level (see Table 1), but with different contexts, making them unique at the document level.

3.3 OntoNotes 5.0 Parallel Corpus
The original OntoNotes comprised various genres of text (news, telephone speech, weblogs, newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with additional annotated information-syntact and predicate argument structure, word sense linked to an ontology and coreference. We extracted the English subsets of broadcast conversations and telephone conversations, and had professional translators translate them into Japanese.

4 Challenges in Translating Conversations
To understand the difficulty of translating conversations, we conducted an analysis of the newly constructed corpus. We especially focus on examples in which context is needed to correctly understand and translate the individual sentences. We analyse the JA→EN translation results of the BSD corpus and EN→JA translation results of the AMI and ON corpora.

4.1 Analysis of the BSD Corpus
In this subsection, we provide an error analysis of the JA→EN MT output using the BSD corpus. We choose to use Google Translate, one of the most powerful publicly available neural machine translation (NMT) systems to produce the translations.

Our primary focus is to understand how many sentences require a context to be properly translated. We randomly sample 10 scenarios (322 sentences) from the corpus, and manually check the translations for critical translation errors, ignoring fluency or minor grammatical mistakes. From the 322 sentences, 12 sentences have errors due to phrase ambiguity that require the context (an example is shown in Figure 1) or the knowledge of real-world situation for a proper translation, and 18 errors due to zero anaphora, which are described in the following section, in the source language (Japanese). Even though they are not that frequent, these types of errors were the only critical errors that we found during our analysis. The more frequent errors were generally minor and did not change the meaning of the translation. Now, we focus on pronoun errors.

---

6 https://translate.google.com/ (May 2019)
4.1.1 Zero Anaphora

As an important preliminary step, we briefly introduce a grammatical phenomenon called zero anaphora. In Japanese, some arguments of verbs are often omitted from phrases when they are obvious from the context. When translating these sentences into English, one often has to identify the referent of the omitted argument and recover it in English, as English does not allow the omission of core arguments (i.e., subject, object). In the example in Figure 2, the subject of the verb 買った is omitted, but in the English translation, a pronoun, for example he, has to be recovered. Note that the subject could be anyone, not necessarily he, depending on the context. The task of identifying the referent of zero anaphora is called zero anaphora resolution, which is one of the most difficult tasks in NLP.

In Japanese, zero anaphora is very pervasive, especially in conversation, as the context of the conversation (either topical or visual) helps identify the referent.

4.1.2 Quantitative Analysis

To estimate how many sentences need zero anaphora resolution in the business conversation corpus, we counted the number of sentences with the personal pronouns (e.g., 彼, 彼女, 私, あなた in Japanese, I, you, he, she in English) in both Japanese and English. As a result, 62% of English sentences contain personal pronouns, whereas only 11% of Japanese sentences do. This means that about 50% of the sentences in the corpus potentially need zero reference resolution they are translated from Japanese into English.

To determine what kinds of zero pronouns are difficult to translate, we again heuristically count the number of translation errors of pronouns for the entire corpus. We counted the number

Context: 「私は総務を担当しています中村です、よろしくお願いします。」
Source: 「ではお揃いのようなので、入社に関しての手続きから行なっていきたいと思います。」
Reference: “It seems we are all here so I would like to start with the process in starting off at our company.”
System Output: ‘Then, it seems like a match, so I would like to start from the procedure for joining.’

Fig. 1 An example of translation error due to phrase ambiguity. The word 「お揃い」 is wrongly translated into a match, which can be resolved by understanding the context, where people are gathered.

太郎は 買った 牛乳を 飲んだ
Taro-SBJ buy-PST milk-OBJ drink-PST
“Taro drank the milk he bought.”

Fig. 2 An example of zero anaphora.
of translated sentences that had pronouns different from their reference sentences. Using this heuristic, we detected 3,653 errors (12% of the whole corpus). The top 10 most frequent errors are shown in Figure 3.

Some errors such as $we \rightarrow I$, $I \rightarrow me$ might not be critical and cannot be regarded as translation errors. However, there are still many critical errors among first-, second-, and third-person pronouns (denoted in boldface in the graph).

Looking at the pronouns that the NMT system produced, we can see the tendency of the system to generate frequent pronouns such as $you$ and $I$. This suggests that the current system tries to compensate source (Japanese) zero pronouns simply by generating frequent target (English) pronouns. When the referent is denoted in relatively infrequent pronouns in the target language, it is difficult to translate correctly. To deal with this problem, we need to develop more sophisticated systems that consider the context.

4.1.3 Qualitative Analysis

This section exemplifies some zero-anaphora translation errors and discusses what kind of information is needed to perform the correct translation.

Translation that Needs World Knowledge and Inference

In Figure 4, the subjects of the verbs are omitted in the source sentence 「(彼は: he) 仕事もあまりしない上に、(彼は: he) 休み、早退ばかりを希望するから」. This causes the NMT system to incorrectly translate the zero pronouns into $I$, although they actually refer to Paul in the previous sentence and thus have to be translated into $he$. Even though there are two persons mentioned in the context sentence, the branch manager and Paul, from the context that Paul is about to be dismissed, it is understandable that Paul is the one who does not work much, takes
frequent days off, and often asks to leave early.

Resolving these zero pronouns, however, is not straightforward, even if one has access to information from the previous sentence. For example, to identify the subject of 「仕事もあまりしない」 (doesn’t work much), one has to know “laziness can lead to being dismissed” and thereby infer that Paul, who is about to be dismissed, is the subject. Existing contextual NMT systems (Voita et al. 2018; Bawden et al. 2018; Maruf et al. 2019) still do not seem to be able to handle this complexity.

**Translation that Needs to Know Who Is Talking**

In Figure 5, again, the subject is omitted in the source sentence 「(君は: you) もう少しの辛抱だよ。」. The NMT system incorrectly translates zero pronouns into I.

It is worth noting that the type of zero pronoun differs from that in Figure 4 in that the referent in Figure 5 does not linguistically appear within the text (called *exophora*), whereas the other does (*endophora*) (Brown and Yule 1983). The referent of the zero pronoun in Figure 5 is the listener of the utterance (you), and it usually does not have another linguistic item (such as the person’s name) that can be referred to. Although some modality expressions and verb types can provide constraints to the possible referents (Nakaïwa and Shirai 1996), the resolution of exophora essentially requires reference to the situation.

**Fig. 4** An example of JA→EN Google Translate output. The words in boldface denote the same referent (Paul).

**Previous Source:**  支店長はボールをクビにするみたいだよ。
**Previous Reference:** It seems like the branch manager will be firing Paul.

**Source:**  仕事もあまりしない上に、休み、早退ばかりを希望するから。
**Reference:** He doesn’t work much, and he takes days off and asks to leave early often.
**System Output:** I do not have much work, and I would like to leave early and leave early.

**Fig. 5** An example of JA→EN Google Translate output. Correct translation requires speaker information.
In this case, the correct translation depends on who is speaking. In the original conversation, the utterance is from Speaker 2 to Speaker 1, and given the context, one can infer that Speaker 2 is speaking to console Speaker 1 and thus, the subject should be you (Speaker 1). However, if the utterance was from Speaker 1, he would then just be complaining about his situation saying, “I just need a bit more patience”. This example emphasises the fact that speaker information is essentially needed to correctly translate some utterances in a conversation.

4.1.4 Notes on EN→JA Translation

So far, our discussion has focused on JA→EN translation. In EN→JA translation, the problem of anaphora usually does not arise; however, the variety of Japanese pronouns poses a problem. One English pronoun usually has many translation candidates in Japanese. For example, the English pronoun you can be translated into the Japanese pronouns, あなた, お前, 君, and many more. They differ in terms of formality, gender, age, and relative social status of the speaker and audience. For the MT system to correctly translate the pronouns, it needs to understand the speaker’s information and the relationship between the speaker and the audience. We leave the resolution of this problem as future work.

4.2 Analysis of AMI and ON Corpora

We manually analyse the AMI and ON corpora in a manner similar to the analysis of the BSD corpus by investigating contextual information requirements for EN→JA MT. We randomly sampled 200 and 100 sentence pairs from the ON and AMI, respectively. In the case of ON, 50% of the pairs were from broadcast conversations, and 50% were from telephone conversations. We translate the sentences using Google Translate and check the translations for errors, ignoring fluency or minor grammatical mistakes.

Unlike the JA→EN results for BSD, where more than 50% of errors were due to zero anaphora, there are mainly two types of causes for errors detected in this analysis: phrase ambiguity and absence of world knowledge. Most of the errors (Table 4) are caused by phrase ambiguity, for which taking context sentences into account can be considered as a possible solution. Conversely, the transcripts in ON broadcast conversations contain a variety of named entities (e.g., Shia, one of the two main branches of Islam) and abbreviations (e.g., CPC, Communist Party of China). To solve this problem, either domain-specific training data or additional mechanisms that account for world knowledge would be required. Even though phrase ambiguity and world knowledge

---

7 https://translate.google.com/ (November 2019)
Machine Translation Experiments

To test our corpus for its intended use case, we trained several configurations of MT systems using only a limited amount of data from the corpus and compared them with systems trained on larger datasets. The total amount of training, development, and evaluation data are summarised in Table 5. WMT is obtained from the WMT20 shared task on MT of news (∼13M parallel sentences mainly consisting of movie subtitles, crawled websites, TED talks and Wikipedia text). It should be noted that WMT is a sentence-level parallel corpus that does not contain document-level information.

### 5.1 Experiment Setup

For the sentence-level systems, we trained the transformer (Vaswani et al. 2017) models with the transformer-base configuration, implemented using the Sockeye (Hieber et al. 2018) library.

To train our context-aware systems, we experimented with two approaches: sentence concatenation (Tiedemann and Scherrer 2017) with source-side factors (Sennrich and Haddow 2016) and context-aware decoder, CADEC (Voita et al. 2019). We used the same toolkit and parameters similar to those in our sentence-level systems for the former and the CADEC toolkit with the default parameters for the latter.
Table 6 Examples of training data source sentences and the respective source-side factors for the concatenated context-aware experiments. The first sentence in the table has no previous context, as it is the first sentence in the respective document. The second sentence has the first sentence as context, followed by the beginning of the sentence tag <bos>. Each C refers to each context token, and each S-source refers to the sentence token to be translated. We consider the beginning of the sentence token (<bos>) as the context.

Sentence Concatenation with Source-side Factors

In the concatenation context-aware MT, the previous sentence from the same document is prepended, followed by a beginning of sentence tag <bos>, to the source sentence. In addition, source-side factors are provided to specify whether a token represents context (“C” in Table 6) or the source sentence to be translated (“S” in Table 6). These factors are converted into embeddings, and each source token embedding is concatenated with the respective factor embedding.

Context-aware Decoder

The CADec toolkit performs translation in two steps. First, a normal transformer model produces an initial translation. This model can be trained with normal sentence-level parallel data without context. After that, the context-aware decoder fixes the translation by incorporating the information from contextual sentences of the source and target language. This decoder is trained using document-level parallel data.

5.2 Results

We evaluate automatic translations using the SacreBLEU tool (Post 2018), which calculates BLEU scores (Papineni et al. 2002) by comparing each automatic translation with a human-produced reference. A higher BLEU score indicates a better translation. The results in the upper part of Table 7 show that decent quality (> ~15 BLEU) MT models can be trained by using only

8 Version string - BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.2.21
Table 7 MT experiment results in BLEU scores on our evaluation dataset. The first four rows use only our corpus (all parts: AMI, ON, BSD) for training, but the training methods change. The last four rows use the baseline training method with different training data setups - only WMT20 data and WMT20 along with our corpus (only the BSD part and AMI and ON parts).

our corpora (AMI, BSD, ON data, and baseline model). For JA→EN the scores slightly improve by training contextual models (concatenated context and concatenated context + source-side factors), which indicates that there are context-dependent sentences in our evaluation set that benefit from the additional information. However, for EN→JA translation, the scores decreased slightly. We investigate this further by performing human evaluation in Section 6.

We were unable to find a clear reason why models trained with CADec underperformed, even our baseline. One possible explanation for the underperformance could be that it uses three context sentences at once for each content sentence. These lines of four consecutive sentences do not overlap with the previous and next four-sentence lines, which effectively shrinks the training data lines to the 1/4th of the original size. Voita et al. (2019), who proposed the CADec, used a 6 million line EN→RU parallel dataset that became 1.5 million in the second training step. In comparison, ours is almost 27x smaller, only 0.22 million in total and 0.055 million in the second step. An obvious way around this would be to use a sliding window of context sentences and generate up to three context sentences for each content sentence; however, we intended to maintain the original approach of the authors for these experiments.

For comparison, we also trained NMT models using WMT data (WMT data row in Table 7). These models reached 21.14 BLEU for EN→JA and 20.43 BLEU for JA→EN on News commentary v15,9 whose domain is close to the WMT data. However, based on our evaluation

---

9 http://www.statmt.org/wmt20/translation-task.html
Source: おっ、きっとお店の後継者になる方ですね。
Reference: Oh, he must be the successor to the store.
Baseline: Oh, I’m sure he will succeed you.
Con.+fact: Oh, I’m sure he will be the successor to the store.

Fig. 6 JA→EN translations of a sentence where the baseline generated an incorrect pronoun, but the concat. + factor system produced a more fitting translation.

data, they underperformed our baselines trained on the proposed corpora. This shows that even with 60x training data, these models struggle to translate conversations. By combining all training data (bottom row in Table 7), the gain over the baselines is only 0.81–1.46 BLEU.

Figure 6 shows one example of a Japanese sentence and its translations by the MT systems. There are no pronouns in the source sentence, but there is the noun 「方」, which should be translated into the English pronoun “he,” specifying the person to be the successor to the store. Both systems manage to translate this part correctly, but the baseline generates an additional pronoun in the end instead of “the store.” We observed many similar situations, where the contextual translation still did not match the reference and was not perfect, but the selection of pronouns had improved.

To verify the usefulness of training on all three parts (AMI, BSD, ON) instead of just BSD, we performed a series of ablation experiments by training the baseline model configuration with different datasets. The two bottom rows of Table 7 show that when training using only BSD the models perform considerably worse than with all three parts combined. However, without BSD in the training data the models performed very poorly on the BSD evaluation set, as expected.

6 Human Evaluation

We translated the evaluation set in both directions using our baseline NMT and performed a two-step human evaluation similar to Voita et al. (2019). Next, we analysed the remaining sentences to determine which sentences really require context.

6.1 Workflow

We used Yahoo! Japan Crowdsourcing\(^{10}\) for human evaluation. We tried to guarantee the evaluation quality using screening questions that were indistinguishable from real questions. Only

\(^{10}\) https://crowdsourcing.yahoo.co.jp/
those who correctly answered all the screening questions were considered valid evaluators. Each sentence was evaluated by five evaluators.

In the first step, evaluators were asked to mark each sentence of the 2051 consecutive sentence pairs\(^{11}\) individually as OK or NG, where OK meant that the translation grasped the general meaning of the original sentence, whereas NG meant that the translation was completely unusable. In the second step, we only used sentence pairs where both sentences were marked OK in the first step by at least three evaluators. We then asked evaluators to mark each sentence pair as OK if the corresponding translations made sense in the context of each other or NG otherwise. We calculated the free-marginal kappa (Randolph 2005) values for the evaluations to measure the agreement between the evaluators. The results (overall agreement, 67%; Free-marginal kappa, 0.34) showed moderate agreement, which is common for crowdsourcing.

### 6.2 Analysis

As a result of the crowdsourcing campaign (Table 8), we had 208 JA→EN sentence pairs and 228 EN→JA sentence pairs marked as NG in the context of each other. Because we used the same method as Voita et al. (2019), we also added their EN→RU results to the table for reference. We employed two linguistic experts to check the translations along with their respective sources and references to determine their ambiguity and the need for additional context. In this step, they were also asked to categorise the ambiguity type. After the final step, 43 JA→EN and nine EN→JA sentence pairs were marked as context-dependent. From the final results, we can conclude the following: 1) using the context may be beneficial for JA→EN translations rather than EN→JA translations, and 2) the crowdsourcing evaluations are not reliable enough to draw direct conclusions even though the evaluation cost is low. From the former point, we can explain

<table>
<thead>
<tr>
<th></th>
<th>JA→EN</th>
<th>EN→JA</th>
<th>EN→RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG in 1st step</td>
<td>669 (32.6%)</td>
<td>892 (43.5%)</td>
<td>211 (10.6%)</td>
</tr>
<tr>
<td>NG in 2nd step</td>
<td>208 (10.1%)</td>
<td>228 (11.1%)</td>
<td>140 (7.0%)</td>
</tr>
<tr>
<td>OK in 2nd step</td>
<td>1,174 (57.2%)</td>
<td>931 (45.4%)</td>
<td>1,649 (82.5%)</td>
</tr>
<tr>
<td>Total sentence pairs</td>
<td>2,051</td>
<td>2,051</td>
<td>2,000</td>
</tr>
</tbody>
</table>

Table 8 Results of the crowdsourcing human evaluation. The results of EN→RU are only for reference from Voita et al. (2019).

\(^{11}\) The number of sentence pairs from each scenario is smaller than the number of sentences by one, because the first sentence of each scenario has no previous context. Therefore, the number of sentence pairs was \(2,120 - 69 = 2,051\).
why the context information is beneficial for JA→EN but not for EN→JA in Table 7. Context-aware models sometimes suffer from noise introduced from context sentences; in other words, the models can generate contents that are not related to the current source sentence but in the context sentences. The effect of noise induction might be larger than the benefit of using context for EN→JA translations.

Thirty-eight JA→EN out of 43 pairs that were marked as context-dependent lack pronouns in the source sentence and did not have enough content to produce an unequivocal translation. One such example is shown in Figure 7. The MT system generated “I” as the subject of the source sentence, which is acceptable without any context. However, considering the previous sentence, using “I” as the subject becomes unnatural. Furthermore, if the speaker information is available (speaker information is available in our corpus) and the MT model can utilise that information, the subject should undoubtedly be “you.” The use of speaker information will be our future work.

The other five JA→EN pairs contain ambiguous words or phrases that can be translated differently depending on the context. For example, 「1組」 can be translated as either “one couple” or “one group.”

Similarly, in EN→JA, Chinese can refer to language (中国語) or food (中華料理), as shown in Figure 8. Our best contextual models still struggle to translate such ambiguities, whereas slightly outperforming sentence-level baselines in handling pronouns.

7 Conclusion

We presented a Japanese–English conversation parallel corpus intended for training and evaluation of MT systems. We describe the corpus in detail and indicate which linguistic phenomena
are challenging for MT. In our evaluation set, we marked examples that can have multiple contrasting translations when tackled on the sentence-level. The release will include the full BSD corpus and Japanese translations of AMI and ON, along with instructions on how to align them. The original source language, speaker, scene, scenario/transcript, and ambiguity types will also be included.

In the future, we plan to model speakers and origin languages in MT, as it can help capture the broader context (Maruf et al. 2018) of the conversations and more precise pronoun translations (Vanmassenhove et al. 2018). We also realise that BLEU is not sufficient for evaluating the translation of zero pronouns. To do this more thoroughly, we would need a dedicated dataset for zero-pronoun evaluation. We intend to use the dataset created by Shimazu et al. (2020) for this task or create our own dataset on the proposed corpus. In addition, we also intend to conduct human evaluations focusing on pronoun translations.

It would also be interesting to conduct deeper analyses on the characteristics of the corpora, such as: a) the effect of bidirectional translations on the BSD corpus; b) linguistic phenomena that exist only in the transcripts (AMI and ON); c) translationese effect at the Japanese side of AMI and ON.

Acknowledgement

This work was supported by the “Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation”, the Commissioned Research of National Institute of Information and Communications Technology (NICT), JAPAN. We want to thank BAOBAB Inc.\textsuperscript{12} for helping us construct the BSD corpus.

\textsuperscript{12} \url{https://baobab-trees.com/}
References


Appendix

A Release Format Example

The three sub-corpora are structured into and are released as json files. Scenarios/transcripts are gathered based on the corpora, and each scenario/transcript is constructed into a single json file, in which sentence pairs are represented as a list of json objects.

Figure 9 shows a sentence pair taken from BSD as an example. Among all fields, “no,” “speaker,” “en_sentence,” “ja_sentence,” “original_language” are common fields available in all the three parts of the corpus. Although most of the field names are self-explanatory, “original_language” is noteworthy here. For AMI and ON, the field is set to English to indicate that both of the original corpora are in English. Conversely, in the case of BSD, it indicates in which language the monolingual scenarios are written in. Additionally, as BSD contains more information such as speaker’s name in Japanese (ja_speaker), scene of the scenario (tag), and title of the scenario (title), we include these fields in the final json files as well.

```
[
  {
    "id": "190315_E001_17",
    "tag": "phone call",
    "title": "Phone: Review spec and scheme",
    "original_language": "en",
    "conversation": [
      {
        "no": 14,
        "en_speaker": "Mr. Sam Lee",
        "ja_speaker": "サム リーさん",
        "en_sentence": "Would you guys consider a different scheme?",
        "ja_sentence": "別の事業案も考慮されませんか？"
      }
    ]
  },
  
  
  ]
```

Fig. 9 An example sentence pair from the Business Scene Dialogue sub-corpus.

---

13 For the BSD corpus, the speaker’s names are taken from real names, but for the AMI and ON corpus, most of the names are anonymised into alphabets as they are not present in the original corpus.
B Human Evaluation Thresholds

For both steps of the human evaluation process, we used a threshold of three concurrent evaluations to count a sentence or sentence pair as “OK.” Table 9 shows how the evaluation results would change for higher and lower thresholds. In this table, the second step represents the case of using a threshold of three for the first step.

<table>
<thead>
<tr>
<th>Minimum OK</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>JA→EN</td>
<td>EN→JA</td>
<td></td>
<td>JA→EN</td>
<td>EN→JA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Step</td>
<td>Both Good</td>
<td>83%</td>
<td>67%</td>
<td>42%</td>
<td>13%</td>
<td>77%</td>
<td>57%</td>
<td>33%</td>
</tr>
<tr>
<td></td>
<td>One/both Bad</td>
<td>17%</td>
<td>33%</td>
<td>58%</td>
<td>87%</td>
<td>23%</td>
<td>43%</td>
<td>67%</td>
</tr>
<tr>
<td>Second Step</td>
<td>Good Pair</td>
<td>64%</td>
<td>57%</td>
<td>42%</td>
<td>19%</td>
<td>53%</td>
<td>45%</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>Bad Pair</td>
<td>3%</td>
<td>10%</td>
<td>25%</td>
<td>48%</td>
<td>4%</td>
<td>12%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 9 Human evaluation results for different thresholds. The one used (3) is marked in bold.

C Human Evaluation Agreement

We calculated the free-marginal multirater kappa values for our human evaluations to measure the agreement between evaluators. The detailed agreement scores are listed in Table 10.

<table>
<thead>
<tr>
<th></th>
<th>JA→EN-1</th>
<th>EN→JA-1</th>
<th>JA→EN-2</th>
<th>EN→JA-2</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall agreement</td>
<td>67.64%</td>
<td>68.74%</td>
<td>65.08%</td>
<td>66.52%</td>
<td>67.00%</td>
</tr>
<tr>
<td>Free-marginal kappa</td>
<td>0.35</td>
<td>0.37</td>
<td>0.30</td>
<td>0.33</td>
<td>0.34</td>
</tr>
<tr>
<td>Fixed-marginal kappa</td>
<td>0.19</td>
<td>0.26</td>
<td>0.09</td>
<td>0.15</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 10 Human evaluation agreement scores for each evaluation step (1 and 2) and the average.

Matīss Rikters: received his B.S., M.S. and Ph.D. degrees from the University of Latvia in 2012, 2014 and 2019 respectively. Currently, he is a postdoctoral researcher at the University of Tokyo. His main field of study is natural language processing and machine translation (MT) in particular. His work has
mainly focused on the improvement of MT for low-resource and morphologically rich languages, and most recently - context-aware MT.

**Ryokan Ri**: is a first-year doctor’s student at the Graduate School of Information Science and Technology, The University of Tokyo. His research interests include natural language processing, especially its multilingual applications.

**Tong Li**: is currently a machine learning engineer at M3 Inc. He received his B.S. in Computer Science and Technology from Jilin University, M.S. in Information Science and Technology from the University of Tokyo in 2017 and 2020, respectively.

**Toshiaki Nakazawa**: received his B.S. in Information and Communication Engineering and M.S. in Information Science and Technology from the University of Tokyo in 2005 and 2007, respectively. He obtained his Ph.D. in Informatics from Kyoto University in 2010. He is currently a specially appointed lecturer at the University of Tokyo. His research interests center on natural language processing, particularly machine translation.

(Received October 28, 2020)
(Revised December 18, 2020)
(Accepted January 18, 2021)