Document-aligned Japanese-English Conversation Parallel Corpus

Matīss Rikters, Ryokan Ri, Tong Li and Toshiaki Nakazawa
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{matiss, li0123, litong, nakazawa}@logos.t.u-tokyo.ac.jp

Abstract

Sentence-level (SL) machine translation (MT) has reached acceptable quality for many high-resourced languages, but not document-level (DL) MT, which is difficult to 1) train with little amount of DL data; and 2) evaluate, as the main methods and data sets focus on SL evaluation. To address the first issue, we present a document-aligned Japanese-English conversation corpus, including balanced, high-quality business conversation data for tuning and testing. As for the second issue, we manually identify the main areas where SL MT fails to produce adequate translations in lack of context. We then create an evaluation set where these phenomena are annotated to alleviate automatic evaluation of DL systems. We train MT models using our corpus to demonstrate how using context leads to improvements.

1 Introduction

The quality of machine translation (MT) for written text and monologue has vastly improved due to the increased amount of available parallel corpora and recent neural network technologies. However, there is much room 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 the pronouns occurs more frequently in spoken language than written language. Recently, context-aware MT models attract attention from many researchers (Tiedemann and Scherrer, 2017; Voita et al., 2019) to solve this kind of problem, however, there are almost no parallel conversation corpora with context information except the rather noisy Open Subtitles corpus (Tiedemann, 2016).

A document and sentence-aligned conversation parallel corpus should be advantageous to push MT research in this field to the next stage. In this paper, we introduce a newly constructed document-aligned (DA) Japanese-English conversation corpus, which contains three sub-corpora: Business Scene Dialogue (BSD (Rikters et al., 2019)), Japanese translation of AMI Meeting Corpus (AMI (McCowan et al., 2005)) and Japanese translation of OntoNotes 5.0 (ON (Weischedel et al., 2011)). The corpus contains multi-person conversations in various situations: business scenes, meetings under specific themes, broadcast conversations and telephone conversations.

We supplement the original BSD part with additional data, increasing its size by almost three times. We also enrich the corpus with speaker information and other useful meta-data, and separate balanced versions of development and evaluation data sets.

2 Related Work

There are many ready-to-use parallel corpora for training MT systems, but most of them are in written languages such as web crawl, patents (Goto et al., 2011), scientific papers (Nakazawa et al., 2016). Even though some parallel corpora are in spoken language, they are mostly 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[1], WAT[2] adopt the written language, monologue or noisy dialogue parallel corpora for their translation tasks. Among them, there is only one clean, dialogue parallel corpus (Salesky et al., 2018) adopted by IWSLT[3] in the conversational speech translation task.

JParaCrawl (Morishita et al., 2019) is a recently announced large English-Japanese parallel corpus built by crawling the web and aligning parallel

[1] http://www.statmt.org/wmt20/
[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/
[3] http://workshop2019.iwslt.org

Proceedings of the 5th Conference on Machine Translation (WMT), pages 639–645
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

sentences. Its size is impressive, but it is composed of noisy web-crawled data and has many duplicate sentences. Compared to our corpus, JParaCrawl does not have meta-information and is not DA.

Voita et al. (2019) evaluate what modern MT systems struggle with when translating from English into Russian and construct new development and evaluation sets based on human evaluation. The sets target linguistic phenomena - dexis, ellipsis and lexical cohesion. The authors also provide code for a context-aware NMT toolkit that improves upon translating these phenomena. In contrast, our development/evaluation sets contain complete documents of consecutive sentences, not broken up into only the sentences requiring context.

3 Corpus Description

Our corpus consists of 3 sub-corpora, each of which originates from different sources - BSD, AMI, and ON. BSD was newly constructed, while AMI and ON are translations of the existing English versions of these corpora. Detailed statistics of the sub-corpora are provided in Tables 1 and 2. BSD consists of the scenes mentioned in Table 1, ON has only two different scenes - broadcast conversation and telephone conversation, and all documents from AMI belong to the meeting scene. There is no particular taxonomy associated with these scenes. Word counts for the English side of the sub-corpora are shown in Table 3. We do not include word counts for the Japanese side since it uses very little spaces and the final word count depends on tokenisation.

3.1 Construction Process

Business Scene Dialogue

This sub-corpus was entirely newly created without using any pre-existing resources. We asked professional scenario writers to write monolingual scenarios (documents), and then asked professional translators to translate the documents. This process was done for both En ↔ Ja directions to ensure a wide range of lexicons and expressions from both languages.

In conversations, the utterances are often very short and vague, therefore it is possible that they should be translated differently depending on the situations where the conversations are taking place. For example, the Japanese expression 「すみません」 can be translated into several English expressions, such as “Excuse me”, “Thank you.” or “I’m sorry.”, depending on context. By using scene information, it is possible to discriminate the translations, which is hard to do with only the contextual sentences. Furthermore, it may be possible to connect scene information to multi-modal MT, i.e., estimating the scene from visual information. Language used in meetings and presentations is often more formal than general chatting or phone calls. This is especially prevalent in Japanese, which has three distinct levels of politeness in the spoken language. Knowing the scene may be useful for adjusting politeness and formality.

AMI Meeting Parallel Corpus

The original AMI Meeting Corpus is a multi-modal dataset containing 100 hours of meeting recordings in English. The parallel version was constructed by asking professional translators to translate utterances from the original corpus into Japanese. Since the original corpus consists of speech transcripts, the English sentences contain a lot of short utterances (*e.g., “Yeah”, “Okay”*) or fillers (*e.g., “Um”*), and these are translated into Japanese as well. Therefore, the AMI sub-corpus contains many duplicates (see Table 6).

OntoNotes 5.0

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

Development and Evaluation Sets

We provide balanced development and evaluation splits from only the BSD sub-corpus as it is the least noisy part. The documents in these sets are balanced in terms of scenes and original languages. The complete statistics are shown in Table 4.

3.2 Analysis

We extend the analysis conducted for BSD (Rikters et al., 2019) to AMI and ON by investigating contextual information requirements for EN→JA

	JA→EN		EN→JA	
Scene	**Doc.**	**Sent.**	**Doc.**	**Sent.**
face-to-face	535	16,481	458	14,858
phone call	279	8,720	256	7,770
general chatting	233	7,674	239	7,372
meeting	224	7,647	265	8,952
training	37	1,379	47	1,549
presentation	17	499	53	1,899
sum	1,325	42,400	1,318	42,400

Table 1: Document (Doc.) and sentence (Sent.) statistics for the full BSD corpus. JA→EN represents documents written in Japanese and translated into English. EN→JA represents the opposite documents.

Set (Scene)	Documents	Sentences	PA	WK
AMI	171	110,483	4	0
ON (BC)	27	14,354	5	3
ON (Tele)	46	14,075	6	0

Table 2: Statistics for translated version of AMI and ON corpora and errors detected in EN→JA MT.

MT. We randomly sample 200 and 100 sentence pairs from ON and AMI respectively. In the case of ON, 50% of the pairs are from BC and 50% are from Tele. We translate the sentences with Google Translate[4] 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 we detected in this analysis - phrase ambiguity (PA) and absence of world knowledge (WK). Most of the errors (Table 2) are caused by PA, for which taking context sentences into account can be considered as a possible solution. On the other hand, the documents in ON-BC 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, either domain-specific training data or additional mechanisms that take WK into account would be required.

3.3 Release and Licensing

The current version of BSD is published on GitHub[5] under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. The English OntoNotes is under the LDC User Agreement for Non-Members and AMI is under Creative Commons Attribution 4.0 license (CC BY 4.0). We plan to release the extended BSD and translations of AMI under the same licenses and are currently negotiating a licensing agreement for the Japanese translations of OntoNotes.

	Word Count
Development	19,229
Evaluation	19,619
BSD	750,167
AMI	977,467
ON	279,709

Table 3: English side word counts for each of the sub-corpora and development/evaluation sets.

4 Machine Translation Experiments

The conversation corpus alone is not big enough to train real-world NMT systems (as demonstrated by Rikters et al. (2019)). However, by increasing the size of the high-quality BSD corpus, we managed to train reasonable NMT systems. The full statistics of our data are shown in Table 6.

4.1 Experiment Setup

For the SL systems, we used Sockeye (Hieber et al., 2017) to train transformer architecture (Vaswani et al., 2017) models with the *transformer-base* parameters until convergence on development data (no improvement on validation perplexity for 10 checkpoints). Each model was trained 3 times on a single Nvidia TITAN V (12GB) GPU. The reported BLEU score results are an average of 3 runs. Training time was about 2 days for models with only our data and about 5 days when using WMT data.

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 use the same toolkit and similar parameters as in our SL systems for the former and the CADec toolkit with the default parameters for the latter. For the concatenation context-aware MT, we experimented with two approaches: 1) prepending the previous sentence from the same document, followed by a beginning of sentence tag <*bos*>, to the source sentence; 2) in addition, providing source side factors to specify if a token represents context or the source sentence.

The source side factors that we used for training were either C or S, representing context and

[4] https://translate.google.com/ (November 2019)

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

	Development				Evaluation			
	JA→EN		EN→JA		JA→EN		EN→JA	
Scene	**Doc.**	**Sent.**	**Doc.**	**Sent.**	**Doc.**	**Sent.**	**Doc.**	**Sent.**
face-to-face	11	319	12	314	12	381	11	345
phone call	6	176	7	185	6	163	7	212
general chatting	7	223	8	248	7	211	8	212
meeting	7	240	7	219	7	228	7	229
training	1	40	1	23	1	38	1	30
presentation	1	31	1	33	1	31	1	40
sum	33	1029	36	1029	34	1052	35	1052

Table 4: Document (Doc.) and sentence (Sent.) statistics for development and evaluation sets.

the actual source sentence respectively. Examples of source sentences with context and factors are shown in Table 5. The first sentence in the table has no previous context, as it is the first one in the respective document. The second sentence has the first one as context, followed by a beginning of sentence tag <*bos*>, and so on.

Source sentences
<bos> はい 、 G 社 お客様 相 談 室 の ケ イ ト です 。
はい 、 G 社 お客様 相 談 室 の ケ イ ト です 。<bos> ご 用 件 は ?
ご 用 件 は ? <bos> もし もし 、 森 と いいます 。
Source side factors
C S S S S S S S S S S S S S S
C C C C C C C C C C C C C C C S S S S S
C C C C C C C S S S S S S S S

Table 5: Examples of training data source sentences and the respective source side factors for the concatenated context-aware experiments.

4.2 Results

The results in Table 7 show that decent quality MT models can be trained by using only our corpus. For JA→EN the scores slightly improve by training contextual models (Concatenated and Concatenated + factors), which indicates that there are context-dependent sentences in our evaluation set that benefit from the additional information. We investigate this further by performing human evaluation in Section 5. We did not find a clear reason why models trained with CADec underperformed even our baseline, but one possible explanation could be that it uses three context sentences at once for each sentence and does not overlap them with the previous and next four-sentence lines, which effectively shrinks the training data down to $\frac{1}{4}$th of the original size.

For comparison, we also trained NMT models on WMT20 data (∼13M parallel sentences, excluding *News Commentary v15*; WMT column in Table 7). For these models, we used *newsdev2020* as development data and *News Commentary v15*[6] as evaluation data since *newstest2020* was not yet available at the time and for Japanese *News Commentary v15* was only 1811 sentences long. These models reached 21.14 BLEU for EN→JA and 20.43 BLEU for JA→EN on *News Commentary v15*, but on our evaluation data they under-performed our baselines. This shows that even with 60x the training data these models struggle to translate conversations. By combining all training data the gain over the baselines is only 0.81 - 1.46 BLEU.

Figure 1 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 didn't match the reference and was not perfect, but the selection of pronouns had improved.

5 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). After that, we analysed the remaining sentences to determine which truly require context.

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

	Total	Unique
Development	2,051	2,012
Evaluation	2,120	2,070
Training	80,629	74,377
AMI	110,483	75,660
ON	28,429	24,335

Table 6: Total vs. unique sentence pairs of training, development and evaluation BSD data; and AMI and OntoNotes sub-corpora.

	JA→EN	EN→JA
WMT	16.29	12.99
WMT+	18.44	15.33
Baseline	16.98	14.52
CADec	15.31	12.55
Concatenated	17.07	14.15
Concatenated + factors	17.24	14.19

Table 7: MT experiment results in BLEU scores. WMT uses only WMT 2020 data and WMT+ uses WMT 2020 along with our corpus for training. The rest use only our corpus for training.

We used Yahoo! Japan Crowdsourcing[7] for the human evaluation. Evaluation quality was guaranteed using screening questions which were indistinguishable from the real questions. Only those who correctly answered all the screening questions were considered valid evaluators. Each sentence was evaluated by 5 different evaluators.

In the first step, evaluators were asked to mark each sentence individually as OK or Not Good (NG), where OK meant that the general meaning of the original sentence was transferred to the translation, whereas NG meant that the translation is completely unusable. In the second step, we used only the consecutive pairs of sentences, which were both marked as OK in the first step by at least three evaluators, and asked evaluators to mark them as OK if the corresponding translations made sense in context of each other. We calculated the Free-Marginal Kappa (Randolph, 2005) values for the evaluations to measure agreement between evaluators. The results (overall agreement - 67%, Free-marginal kappa - 0.34) show moderate agreement, which is common for crowdsourcing.

5.1 Analysis

As a result of the crowdsourcing campaign (Table 8) we had 228 EN→JA sentence pairs and 208 JA→EN sentence pairs marked as NG in context of each other. We employed two linguistic experts to check the translations along with their respective sources and references to determine their ambiguity and need for additional context. For this step they were also asked to categorise the ambiguity type.

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.

Figure 1: JA→EN translations of a sentence where the baseline generated an incorrect pronoun, but the concat. + factors system produced a more fitting translation.

Previous Source:	What kind of food should we choose?
Previous Reference:	どういうジャンルにしますか？
Previous MT:	どんな食べ物を選ぶべきか。
Source:	How about **Chinese**?
Reference:	中華料理はどう？
MT:	中国語はどうですか？

Figure 2: EN→JA MT output where *Chinese* is translated into "中国語" (Chinese language) instead of "中華料理" (Chinese food).

After the final step 9 EN→JA and 43 JA→EN sentence pairs were marked as context-dependent. 38 JA→EN pairs lack pronouns in the source sentence and do not have enough content to produce an unequivocal translation. The other 5 JA→EN pairs contain ambiguous words or phrases, which 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 2. Our best contextual models still struggle to translate such ambiguities, while slightly outperforming SL baselines in handling pronouns.

Figure 3 shows example mistranslations of pronouns, where they are omitted (as is often done in the spoken language) on the Japanese side, but expected in the English translation. The contextual MT model does get some of the pronouns right in the first sentence, but perhaps requires longer context for the second one.

6 Conclusion

We presented a document-aligned parallel corpus of English-Japanese conversations 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

[7] https://crowdsourcing.yahoo.co.jp/

EN→RU		EN→JA		JA→EN	
2000		2051		2051	
NG	OK	NG	OK	NG	OK
140	1649	228	931	208	1174
4%	41%	11%	45%	10%	57%

Table 8: Results of the second step of the crowdsourcing human evaluation compared to EN→RU (Voita et al., 2019). The first row shows sentence pair totals and the last two rows show sentence pairs, where both sentences were marked as "good" individually, evaluated in context of each other as either good or bad pairs.

Prev. Source:	いつ 返事 くれる と 言っ て た？
Prev. Reference:	Did they say when they will get back to you?
Prev. Base.:	when did you say you' d answer me?
Prev. Conc.+f.:	When did they say they will reply?
Source:	来週 早々 に は 、 と 言っ て まし た 。
Reference:	They said early next week.
Base.:	He told me early next week.
Conc.+f.:	I said it early next week .

Figure 3: JA→EN MT output by baseline (Base.) and concatenated context + factored (Conc.+f.) models of sentences with no pronouns in the source and expected pronouns in the translation.

evaluation set we marked examples, which 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, document, ambiguity type will also be included.

In the future we plan to model speakers and origin languages in MT, as it can help capture broader context (Maruf et al., 2018) and more precise pronoun translations (Vanmassenhove et al., 2018). We are also interested in experimenting with modelling the scene information within the training data to produce more appropriate translations for each of the politeness settings.

Acknowledgements

This work was supported by "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.

References

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin Tsou. 2011. Overview of the patent machine translation task at the ntcir-9 workshop. In *Proc. of NTCIR-9 Workshop Meeting*, pages 559–578.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *ArXiv e-prints*.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Belgium, Brussels. Association for Computational Linguistics.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2019. Jparacrawl: A large scale web-based english-japanese parallel corpus. *arXiv preprint arXiv:1911.10668*.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English Subtitle Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater κfree): An alternative to fleiss' fixed-marginal multirater kappa. In *Presented at the*

Joensuu Learning and Instruction Symposium, volume 2005.

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.

Elizabeth Salesky, Susanne Burger, Jan Niehues, and Alex Waibel. 2018. Towards fluent translations from disfluent speech. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, Athens, Greece.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A Large Training Corpus for Enhanced Processing*, chapter 1. Handbook of Natural Language Processing and Machine Translation. Springer.