Appendix

A Corpus Example

The utterances in conversations 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. What is unique for our corpus is that each scenario is annotated with scene information, as shown in the Table 1.

Japanese		English	
Speaker	**Content**	**Speaker**	**Content**
佐々木	もしもし、A社の佐々木です。	Sasaki	Hello, this is Sasaki from A Corporations.
本田	こんにちは。	Honda	Good afternoon.
本田	御社の製品についてちょっとお聞きしたいのですが。	Honda	I have a few questions about your products.
佐々木	はい、どうぞ。	Sasaki	Yes, please.
本田	カナダ旅行に持っていくワイファイ機器をレンタルしようと思っていました。	Honda	I was thinking of renting one of your Wi-Fi devices for my trip to Canada.
本田	アルバータ州でも使えるものありますか？	Honda	Do you have anything that will work in Alberta?
佐々木	はい、あります。	Sasaki	Yes, we do.
...	...	...	...

Table 1: An example of the Japanese-English business conversation parallel corpus. Scene: telephone inquiry about products.

B Release Format Example

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

Figure 1 shows a sentence pair taken from BSD as an example. Among all of the fields, “no”, “speaker”[1], “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 worth to note here. For AMI and ON, the field is set to English to indicate that both of the original corpora are in English. On the other hand, in the case of BSD, it indicates in which language the monolingual scenarios are written in. Additionally, since 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.


```
[
    {
        "no": 14,
        "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すか？",
        "original_language": "en",
        "tag": "phone call",
        "title": "Phone: Review spec and scheme"
    },
]
```


Figure 1: Example sentence pair from the Business Scene Dialogue sub-corpus.

[1]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 anonymized into alphabets as they are not present in the original corpus.

C Human Evaluation Thresholds

For both steps of the human evaluation process, we used a threshold of three agreeing evaluations to count a sentence or sentence pair as "OK". Table 2 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 3 for the first step.

	Minimum OK							
	2	**3**	**4**	**5**	**2**	**3**	**4**	**5**
	EN→JA				**JA→EN**			
	First Step							
Both Good	77%	**57%**	33%	12%	83%	**67%**	42%	13%
One/both Bad	23%	**43%**	67%	88%	17%	**33%**	58%	87%
	Second Step							
Good Pair	53%	**45%**	35%	18%	64%	**57%**	42%	19%
Bad Pair	4%	**12%**	22%	39%	3%	**10%**	25%	48%

Table 2: Human evaluation results for different thresholds. The one used (3) is marked in bold.

D Human Evaluation Agreement

We calculated the Free-Marginal Multirater Kappa values for our human evaluations to measure agreement between evaluators. The detailed agreement scores are shown in Table 3.

	EN→JA-1	**JA→EN-1**	**EN→JA-2**	**JA→EN-2**	**AVG**
Overall agreement	68.74%	67.64%	66.52%	65.08%	67.00%
Free-marginal kappa	0.37	0.35	0.33	0.30	0.34
Fixed-marginal kappa	0.26	0.19	0.15	0.09	0.17

Table 3: Human evaluation agreement scores for each evaluation step (1 and 2) and the average.