DIRECT: Direct and Indirect Responses in Conversational Text Corpus

Junya Takayama[†], **Tomoyuki Kajiwara**[‡], and **Yuki Arase**[†]
[†] Graduate School of Information Science and Technology, Osaka University,
[‡] Graduate School of Science and Engineering, Ehime University,
[†] {takayama.junya, arase}@ist.osaka-u.ac.jp
[‡] kajiwara@cs.ehime-u.ac.jp

Abstract

We create a large-scale dialogue corpus that provides pragmatic paraphrases to advance technology for understanding the underlying intentions of users. While neural conversation models acquire the ability to generate fluent responses through training on a dialogue corpus, previous corpora have mainly focused on the literal meanings of utterances. However, in reality, people do not always present their intentions directly. For example, if a person said to the operator of a reservation service 'I don't have enough budget.', they, in fact, mean 'please find a cheaper option for me.' Our corpus provides a total of $71,498$ indirect–direct utterance pairs accompanied by a multi-turn dialogue history extracted from the MultiWoZ dataset. In addition, we propose three tasks to benchmark the ability of models to recognize and generate indirect and direct utterances. We also investigated the performance of state-of-the-art pre-trained models as baselines.

1 Introduction

We create a large-scale dialogue corpus that discloses users' hidden intentions to advance techniques for natural language understanding in dialogue systems. Neural conversation models have been able to generate high-quality responses (Zhao et al., 2020; Zhang et al., 2020) and achieve dialogue state tracking (Hosseini-Asl et al., 2020; Lin et al., 2020). These previous studies have been based on the *literal* meanings of user utterances. Little attention has been given to the implied intention of the utterances considered.

However, during conversation, humans often respond to others with indirect expressions, without directly telling them their requests or intentions (Searle, 1979; Brown et al., 1987). When humans receive an indirect response, they infer the intention implied in the response based on context, such as dialogue history. For example, in the example of operator–user dialogue in Figure 1, the user responds with 'I don't have enough budget' to the operator's utterance of 'Would you like to make a reservation for this restaurant?' If the operator only considers the literal meaning, they would repeat the question. However, based on the dialogue history, the operator would infer that the user wants a cheaper restaurant and thus suggest an option to satisfy the user's preference. Our experiments revealed that even a state-of-the-art dialogue system (Yang et al., 2021) degrades the quality of response generation for indirect utterances. Such a pair of user utterances and hidden intentions is categorized into the class of *pragmatic* paraphrases, which emerge in conversations depending on the context. To realize dialogue systems for communicating with users at the human level, the systems should process the pragmatic paraphrases to address the true intentions of the user.

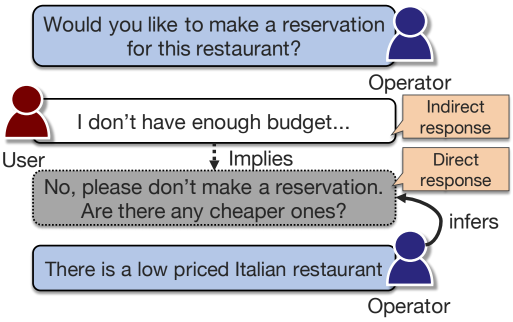

Would you like to make a reservation for this restaurant?
Operator
I don't have enough budget...
User
Indirect response
Implies
No, please don't make a reservation. Are there any cheaper ones?
Direct response
infers
There is a low priced Italian restaurant
Operator

Figure 1: An example of an indirect or direct response in a conversation. Although their literal meanings are different, they can be paraphrased in this dialogue.

In this study, we release[1] a corpus of direct and indirect responses in conversational text, *DIRECT*, which contains $71,498$ pairs of indirect and direct responses. We expand the commonly used dialogue corpus of MultiWoZ (Eric et al., 2020), a multi-domain and multi-turn task-oriented dialogue

[1] https://github.com/junya-takayama/DIRECT

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1980–1989
November 7–11, 2021. ©2021 Association for Computational Linguistics

corpus. The MultiWoZ corpus is created using the Wizard-of-Oz method, in which the user and system speak alternately. For each *user*'s utterance, we use crowdsourcing to collect 'an utterance that is more indirect than the original utterance' and 'an utterance that is more direct than the original utterance'. Hence, *DIRECT* provides triples of paraphrases: original utterances, indirect utterances, and direct utterances.

We designed three benchmark tasks using this corpus to evaluate the model's ability to recognize and generate pragmatic paraphrases. As baselines, we investigated the performance of state-of-the-art pre-trained models, BERT (Devlin et al., 2019) and BART (Lewis et al., 2020), for benchmark tasks.

2 Related Work

Paraphrases have been applied in a dialogue system's research in the context of data augmentation (Hou et al., 2018; Gao et al., 2020). Despite its importance in understanding users' intentions, the pragmatic paraphrases have been overlooked. Only a few recent studies have focused on pragmatic paraphrases to advance the understanding of users' intentions. Pragst and Ultes (2018) proposed a rule-based approach to automatically construct a corpus consisting of pairs of indirect and direct utterances. They demonstrated that the neural conversation model could accurately extract utterances with opposing directness. Because of their rule-based approach, patterns of indirect/direct utterances in their corpus are limited. Louis et al. (2020) used crowdsourcing to build a corpus comprising indirect answers to Yes/No questions, annotating whether the answers were Yes or No. This corpus provides natural answers written by crowdsourcing workers; however, it is limited to context-free Yes/No questions. In contrast to these studies, *DIRECT* provides natural utterances written by humans with rich dialogue histories. Furthermore, it covers various types of utterances.

While there are several paraphrase corpora (Dolan and Brockett, 2005; Lan et al., 2017), all have focused on context-free paraphrases. Hence, none provide pragmatic paraphrases that emerge in contexts. Corpora for natural language inference are also relevant to our study (Giampiccolo et al., 2007; Marelli et al., 2014; Bowman et al., 2015). Similar to the paraphrase corpora, they do not provide contexts. This means that these corpora rely on world knowledge to determine whether a text entails a hypothesis. In contrast, context is a crucial element in determining paraphrasal relationships in pragmatic paraphrases. Our *DIRECT* is the first corpus that provides large-scale pragmatic paraphrases. It would be a valuable resource also for research on paraphrase identification and generation to make a step forward from literal paraphrases.

3 DIRECT Corpus

A pragmatic paraphrase is a pair of texts that have equivalent outcomes in a given context, which frequently emerge in conversations. Expanding a dialogue corpus is a promising approach for building a corpus that collects such pragmatic paraphrases as such a corpus is conversational by nature and often provides conversation histories. Specifically, we employed MultiWoZ2.1 (Budzianowski et al., 2018; Eric et al., 2020) and collected pragmatic paraphrases using crowdsourcing.

We describe how we collected pragmatic paraphrases in Section 3.1 with careful quality control as described in Section 3.2. Section 3.3 describes the statistics of the collected corpus. Section 3.4 presents a comparative analysis between our corpus and existing paraphrase corpora using conventional paraphrase identification models.

3.1 Direct and Indirect Response Collection

MultiWoZ is a multi-domain, task-oriented dialogue corpus annotated with dialogue act tags and dialogue states, comprising $10,438$ dialogues. Each dialogue involves alternate utterances by a user and system; the total number of utterances is $71,524$.

We used Amazon Mechanical Turk[2], a crowdsourcing service, to expand MultiWoZ with pragmatic paraphrases. The workers first received instructions, as presented in Table 1, and some examples of the task. Then, the workers were shown dialogue histories extracted from MultiWoZ, as illustrated in Figure 2.[3] Based on the given conversation histories, the workers input indirect and direct responses that have the same intent as the specified user response in the dialogue (written in red in Figure 2) into the input forms at the bottom.

[2] https://www.mturk.com/

[3] The original MultiWoZ data contains dialogues between a person acting as the 'system' and another acting as the 'user.' We presented the former as the 'operator' in our user interfaces to prevent the workers from misunderstanding the corresponding utterances that were automatically generated.

Instructions
Read the following dialogue between the USER and the OPERATOR, please rephrase the USER's response written in red letters into two different types of speech, following the instructions below.
Type-1 (Direct) : a more direct response that expresses the same intention as the original response **Type-2 (Indirect) :** a more indirect but natural response that expresses the same intention as the original response
'Indirect response' means, for example, a response to a Yes/No question that does not contain a 'Yes' or 'No', or a response that does not directly refer to the action you want the other person to do or your desire. If you have trouble rephrasing, click the 'Hints' button. You can see the goals that 'USER' must achieve in that interaction.

Table 1: Instructions for workers

We assumed that workers should be able to develop indirect and direct responses based only on the dialogue histories. If they did not understand the intent of the utterance that they were required to paraphrase, we provided an option to refer to the goal of the user as 'Hints' (upper part of Figure 2). Such goals were extracted from the MultiWoZ.

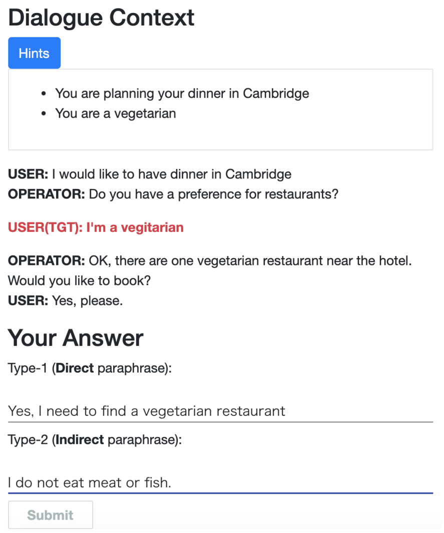
We targeted the utterances of the 'user' in MultiWoz for paraphrasing because users primarily express their needs and preferences. We assumed the average time per post to be 1 minute and set the average reward at 0.12 USD (7.2 USD per hour).

As a result, we collected $71,498$ indirect–direct pairs. We divided the corpus into training and test data in the same manner as in the settings of MultiWoZ. Note that our corpus is a parallel corpus, comprising indirect and direct responses, but it can also be used with the original MultiWoZ responses, *i.e.*, triples of indirect, original, and direct responses are also extractable.

Examples Table 2 presents the examples of the collected pragmatic paraphrases. In the upper example, the user asks for a restaurant in a moderate price range. The indirect response is 'I don't want to overspend but remember its also vacation,' which requires an understanding that 'its also vacation' is a paraphrase of 'not too cheap', as explicitly stated in the direct response in this context. In the lower example, the phrase 'Do you know of any in town?' in the indirect response paraphrases 'Can you find me a guesthouse...?' in this context.

3.2 Quality Control

We carefully created the *DIRECT* corpus to collect high-quality pragmatic paraphrases by prescreening workers. We also conducted a quality assessment.



Dialogue Context

- You are planning your dinner in Cambridge
- You are a vegetarian

USER: I would like to have dinner in Cambridge
OPERATOR: Do you have a preference for restaurants?

USER(TGT): I'm a vegitarian

OPERATOR: OK, there are one vegetarian restaurant near the hotel. Would you like to book?
USER: Yes, please.

Your Answer

Type-1 (**Direct** paraphrase):

Yes, I need to find a vegetarian restaurant

Type-2 (**Indirect** paraphrase):

I do not eat meat or fish.

Figure 2: User interface shown to crowdsourcing workers to generate indirect and direct paraphrases

Worker Selection Prior to formal data collection, we carefully selected crowd workers to avoid trivial paraphrases by replacing or shuffling some words. Specifically, we posted a pilot consisting of 2 tasks. We automatically rejected workers whose average word-level Jaccard index between indirect and direct responses exceeded 0.75. We also manually observed sampled paraphrases. We then chose workers to ask for actual tasks that passed these automatic and manual quality assessments. In total, we obtained 536 workers to exclusively complete the tasks.

Quality Assessment After completing paraphrase collection, we used the same crowd workers to assess the quality of the collected pragmatic paraphrases for $7,372$ dialogues from the test set. We showed the workers utterances for assessment with their dialogue histories. The paraphrased utterances, presented as Response-A and Response-B, were also shown to the worker, of which indirect or direct labels were closed. Using a binary label, the workers first judge whether paraphrased utterances have the same intention as an original utterance. The workers then determined whether Response-A or Response-B was more direct. If the workers could not make a decision, they were allowed to choose a 'no difference' label.

We assumed the average time per post to be 30 seconds, and set the reward at 0.06 USD (7.2 USD per hour). Five workers were assigned to each

speaker	utterance
	(a user is looking for a restaurant.)
SYSTEM	Would you like to pick a different type of food?
USER	Yes, what about British food please.
SYSTEM	What price range are you comfortable with?
USER (original)	Something in the moderate price range would be good.
USER (indirect)	I dont want to overspend but remember its also vacation.
USER (direct)	Can you choose something that is not too expensive and not too cheap.
SYSTEM	Do you have a preference as to what area of town you dine in?

speaker	utterance
USER	I need a place to stay in the north
SYSTEM	OK im seeing alot of choices in hotels is there anything else You need in the hotel that would help narrow it down
USER (original)	I'd really like to stay in a guesthouse. I heard the ones in Cambridge are very nice.
USER (indirect)	I am thinking of staying in a guesthouse. Do you know of any in town?
USER (direct)	Can you find me a guesthouse in Cambridge?
SYSTEM	How about the Acorn Guesthouse? It is rated 4 stars and is in the moderate price range.

Table 2: Examples of *DIRECT* corpus. 'USER (indirect)' and 'USER (direct)' are the responses created by the crowd worker based on 'USER (original)'.

Metric	Ratio [%]
Intention-accuracy (Indirect)	95.0
Intention-accuracy (Direct)	99.7
Directness-accuracy	81.4

Table 3: Quality assessment results

paraphrase and the final label was decided via majority voting. Note that in this assessment task, a worker was assigned paraphrases generated by another worker to avoid self-evaluation. The assessment results are listed in Table 3. Intention-accuracy is the percentage of collected responses that were judged to have the same intention as the original response. Intention-accuracy for both indirect and direct paraphrases is 95.0% and 99.7%, respectively, indicating that the collected sentences preserve the same intent of the original utterances. The intention-accuracy of indirect responses was 4.7% lower than direct responses. This is expected because these utterances indirectly represent users' intentions, which makes the utterance more or less ambiguous.

Directness-accuracy is the percentage of direct responses judged as 'direct' by the worker. The accuracy was as high as 81.4%. The *DIRECT* corpus also provides these assessment labels.

Metric	Value [words]
Vocabulary size (Indirect)	6, 273
Vocabulary size (Direct)	4, 664
Length (Indirect)	15.59
Length (Direct)	12.38
Keep (Indirect-to-Direct)	5.33
Add (Indirect-to-Direct)	7.04
Delete (Indirect-to-Direct)	10.26

Table 4: Statistics of collected paraphrases

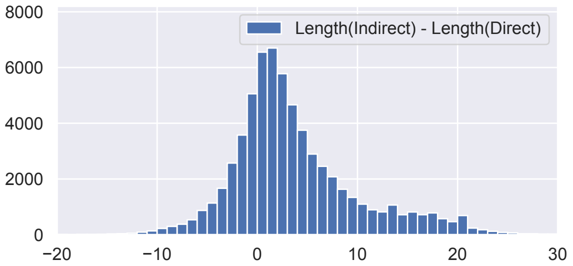


Figure 3: Distribution of differences in sentence lengths between the indirect and direct responses

3.3 Statistical Analysis

We reveal the characteristics of pragmatic paraphrases in the *DIRECT* corpus using case-insensitive token-level analyses. Table 4 presents the word-based statistics on our corpus (except the test data).[4] First, the vocabulary size of indirect responses was much larger than that of direct responses. This implies that even for utterances with the same intent, there are more diverse expressions in the indirect responses than in the direct responses. The average number of words in utterances was 15.59 for indirect utterances and 12.38 for direct utterances. Wilcoxon's test (Wilcoxon, 1945) confirmed that the difference in the average number of words in utterances was statistically significant at the level of 0.1%. Figure 3 shows the histogram of differences in lengths between indirect and direct responses, where the distribution spreads to both positive and negative ranges. This implies that simply shortening a sentence does not necessarily make an utterance more direct.

Next, we investigate the number of words that need to be replaced to transform an indirect response into a direct one. We computed three metrics of 'Keep', 'Delete', and 'Add'. 'Keep' is the average number of words kept when rewriting indirect responses to direct, 'Add' is the number of words that need to be added and 'Delete' is the

[4] For tokenization, we used word_punct_tokenize() method of the nltk (`https://www.nltk.org/`) library.

Indirect responses		Direct responses	
trigram	freq	trigram	freq
i want to	2387	**find me a**	2617
i need to	2223	i want to	2442
would like to	1969	**can you find**	1971
i would like	1903	**please find me**	1924
is there any	1762	all i needed	1807
that would be	1588	thanks for the	1643
you help me	1584	in the centre	1628
thanks a lot	1407	i need to	1623
in the centre	1402	for the help	1539
i think that	1312	**you find me**	1531
i think i	1270	that's all i	1403
a place to	1246	can you get	1376
you have been	1242	give me the	1358
would be swell	1056	i need a	1206
such a great	1042	you get me	1200
i think you	1027	i would like	1145
you have done	1011	please give me	1097
a great help	981	get me a	1094
have been such	979	**book it for**	1086
been such a	979	**the reference number**	1060

Table 5: Top 20 frequent trigrams

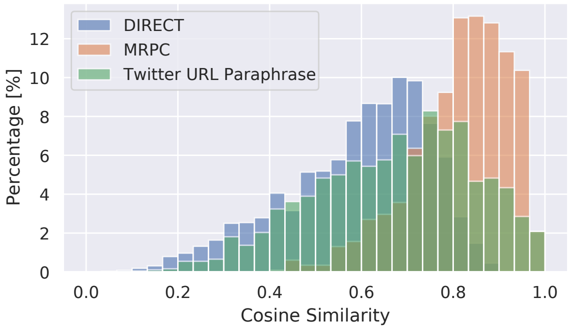


Figure 4: Histograms of cosine similarities of sentence embedding (encoded by Sentence-BERT) for paraphrase pairs.

number of words that need to be deleted. Table 4 demonstrates that 'Keep' is smaller than 'Add' and 'Delete.' This indicates that more than half of the words need to be replaced to transfer an indirect response into a direct response.

Finally, Table 5 presents the top 20 most frequent trigrams that appear in indirect and direct responses. Indirect and direct responses use distinctive expressions. Frequent trigrams of direct responses contain verbs that directly convey what the user wants an operator to do, such as 'book' and 'find', as well as phrases that refer to specific objects, such as 'the reference number'. On the contrary, trigrams of indirect responses contain phrases of 'is there any' and 'I think that', which do not appear in the counterpart.

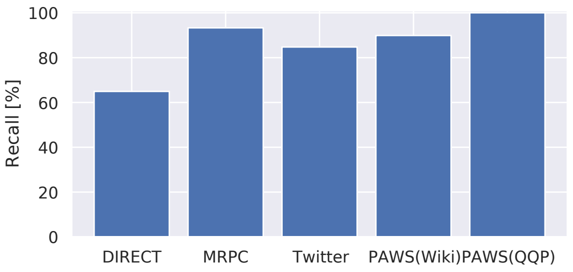


Figure 5: Percentages of paraphrased pairs recognized as paraphrases by BERT fine-tuned on MRPC and Twitter URL paraphrase corpus.

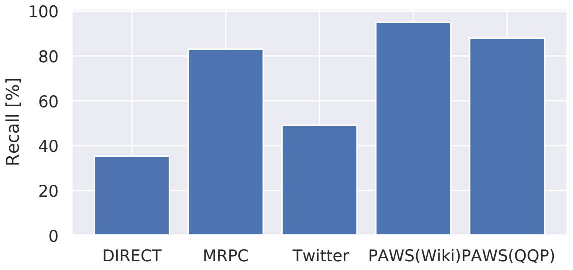


Figure 6: Percentages of paraphrased pairs recognized as paraphrases by BERT fine-tuned on PAWS.

3.4 Model-based Analysis

In this section, we investigate how the *DIRECT* corpus differs from existing paraphrase corpora using state-of-the-art paraphrase identification models.

First, we compute the cosine similarity between paraphrase pairs in *DIRECT*, MRPC (Dolan and Brockett, 2005), and Twitter URL Paraphrase corpus (Lan et al., 2017) using Sentence-BERT[5] (Reimers and Gurevych, 2019). Figure 4 shows the histograms, which confirms *DIRECT* provides more paraphrase pairs with lower cosine similarities than the MRPC and Twitter URL Paraphrase corpus. Sentence-BERT is expected to address the literal meaning of a sentence through its pre-training via STSBenchmark (Cer et al., 2017). The large volume of paraphrases with lower similarities confirms that *DIRECT* provides paraphrases beyond literal similarity.

Next, we investigate whether a paraphrase identification model trained on existing paraphrase corpora transfers to *DIRECT*. Specifically, we calculated the percentage of paraphrase sentence pairs that are recognized as paraphrases using the paraphrase identification model. Figures 5 and 6 show the results where BERT (Devlin et al., 2019) was

[5] We used the 'stsb-roberta-base' model available at https://www.sbert.net/.

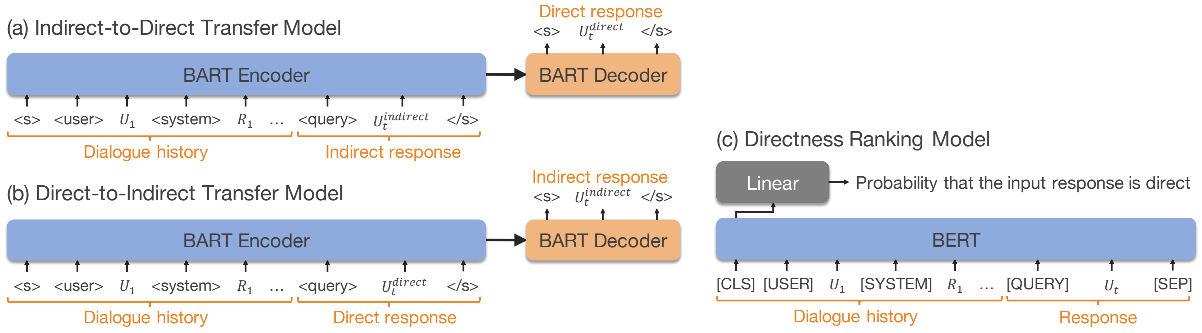

(a) Indirect-to-Direct Transfer Model
Direct response
<s> U_t^{direct} </s>
BART Encoder
BART Decoder
<s> <user> U_1 <system> R_1 … <query> $U_t^{indirect}$ </s>
Dialogue history
Indirect response
(b) Direct-to-Indirect Transfer Model
Indirect response
<s> $U_t^{indirect}$ </s>
BART Encoder
BART Decoder
<s> <user> U_1 <system> R_1 … <query> U_t^{direct} </s>
Dialogue history
Direct response
(c) Directness Ranking Model
Linear
Probability that the input response is direct
BERT
[CLS] [USER] U_1 [SYSTEM] R_1 … [QUERY] U_t [SEP]
Dialogue history
Response

Figure 7: Architectures of baseline models

fine-tuned[6] using MRPC and Twitter URL Paraphrase corpus and Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019), respectively. The PAWS corpus is created to focus on syntax in paraphrases; hence, it has the highest word overlap rate. The results show that *DIRECT* has the lowest percentage of pairs recognized as paraphrases for both BERT models: only 64.9% and 35.3%, respectively. This indicates that pragmatic paraphrases in *DIRECT* are exclusive to existing paraphrase corpora and are difficult to recognize via models trained on literal paraphrases.

4 Benchmark Tasks

We designed three tasks using the *DIRECT* corpus to benchmark the models' ability to handle pragmatic paraphrases. Specifically, we design an indirect-to-direct transfer task (Section 4.1), direct-to-indirect transfer task (Section 4.2), and directness prediction task (Section 4.3). We also evaluated state-of-the-art pre-trained models for these tasks as baselines.

4.1 Indirect-to-direct Transfer Task

Task Description and Motivation Indirect-to-direct transfer is the task of transforming an indirect response into a direct response while preserving its intent under the context, *i.e.*, the dialogue history. This task allows the ability of a certain model to accurately interpret the intent of an indirect response to be evaluated. A possible application of this task is the pre-editing of utterances for task-oriented dialogue systems. By transforming the user's indirect utterances into direct utterances that are easier to interpret before inputting them into the model, the response generation quality is expected to improve.

Baselines We employed BART (Lewis et al., 2020) as the baseline model for this benchmark. The architecture of the baseline model is shown in Figure 7 (a). We added new special tokens '<user>', '<system>', and '<query>' such that the model can distinguish between utterances in the dialogue history and the response to transform. We first added a '<query>' tag to the beginning of the indirect response to the transformation. Then, for the dialogue history, we added '<user>' and '<system>' at the beginning of user utterances and system utterances, respectively. These utterances are concatenated in the order of appearance in the dialogue history and input into the BART encoder. We fine-tuned the model using cross-entropy loss.

For implementation, we used the transformers (Wolf et al., 2020) library. The pre-trained model we used was 'facebook/bart-base'.[7] We used the AdamW (Loshchilov and Hutter, 2019) optimizer for training, with a learning rate of 2e-5.[8] The batch size was 8, owing to the GPU memory size.[9] We randomly sampled $2,000$ dialogues from the training data as a validation set and used the rest for training. Through 30 epochs of training, the model with the lowest validation loss was used to evaluate the test data.

We also constructed a model that disregards dialogue history (BART without history) to investigate the effects of context. In addition, we trained a transformer model (Transformer w/ history) from scratch on only the *DIRECT* corpus to investigate the effects of pre-training. The transformer model has the same architecture as the BART, comprising six self-attention layers of the encoder and decoder.

[6] We used 'bert-base-cased' model in transformers library https://huggingface.co/transformers/ (version 3.5.1).

[7] https://huggingface.co/facebook/bart-base

[8] We chose the value that gives the lowest validation loss.

[9] GeForce RTX 2080 Ti, 11 GB Memory

Model	BLEU	Perplexity
Transformer w/ history	25.23	2.66
BART w/o history	32.51	2.15
BART w/ history	**33.77**	2.16

Table 6: Evaluation results of the indirect-to-direct task

USER	Hi, I'm looking for a place to eat some indian food.
SYSTEM	Do you have a price range in mind?
Indirect response	I want it to be the best place in town.
BART	I want it to be an expensive place.
- w/o history	Can you find me something expensive?
transformer	Can you find me an attraction in town?
reference	I was hoping for a much more expensive place if possible.
	(omitted)
SYSTEM	Yes, The Cambridge Belfry is a cheap hotel in the West.
USER	Do they have free internet?
SYSTEM	Yes, they have internet would you like me to book it for you?
indirect response	i dont think so on the booking. address and phone number though.
BART	dont need to book it. address and phone number for them though.
- w/o history	address and phone number isnt needed.
transformer	no...just give me the address and phone number.
reference	address and phone number is all i need right now.

Table 7: Examples of generated direct responses in the indirect-to-direct transfer task.

Results and Discussion Table 6 presents the BLEU score (Papineni et al., 2002) and Perplexity of each model. For the BART-based models, the model using dialogue history has a higher BLEU score, as expected, because pragmatic paraphrases are context-dependent. Comparing BART and transformer with dialogue history, the former largely outperformed the latter. This result confirms that pre-training is also crucial in this task.

Examples of generated direct responses are presented in Table 7. In the upper example, the BART model successfully interprets the phrase 'the best place' in the indirect response as an expression of the price range, while the transformer without pre-training failed in this interpretation. The context is particularly important in the indirect response in the lower example. In fact, the BART w/o history model generated a sentence with the opposite intent to the reference. Conversely, the two models that use dialogue history generate sentences with the same intent as the reference.

Model	BLEU	Perplexity
Transformer w/ history	19.84	3.06
BART w/o history	27.12	2.39
BART w/ history	26.52	2.34

Table 8: Evaluation results of the direct-to-indirect task

	(omitted)
USER	Thanks! I'm also looking for the Curry Prince restaurant, do you know where that is?
SYSTEM	Yes, it is located at 451 Newmarket Road in Fen Ditton. Can I book a table for you?
direct response	Let me know around where that place is if you dont mind.
BART	I have no idea where anything is at in this town.
- w/o history	I have no idea where that is even at.
Transformer	I just want to make sure i dont have to worry about them leaving a ticket.
Reference	Was hoping you could tell me what direction to head in.

Table 9: An Example of generated indirect responses in the Direct-to-Indirect transfer task.

4.2 Direct-to-Indirect Transfer task

Task Description and Motivation Direct-to-indirect transfer is a task, in contrast to the previous one, that transforms a direct response into an indirect response while preserving its intent. Miehle et al. (2018) have shown that there are approximately the same number of users of dialogue systems who prefer indirect responses as users who prefer direct responses. In addition, indirectly expressing requests to others is a polite strategy to save their face (Brown et al., 1987). Hence, for dialogue systems to have smooth and polite communication with humans, a technology to rephrase a direct response into an indirect one is also desired.

Baselines Similar to the setup in the indirect-to-direct task, we used the BART model that takes a dialogue history as input as a baseline. The model architecture is shown in Figure 7 (b). We input the dialogue history and direct utterance into BART in the same manner as described in Section 4.1. We also constructed a BART model that disregards dialogue history, as well as a transformer trained on the *DIRECT* corpus from scratch. The hyperparameters and training settings are the same as those in the indirect-to-direct task.

Results and Analysis Table 8 shows the BLEU and Perplexity. The fine-tuned BART models achieved higher BLEU scores and lower perplexity than the transformer without pre-training, which again confirms the effectiveness of pre-training.

Overall, the performances of all models were lower in this task than in the indirect-to-direct transfer task. Moreover, dialogue history did not improve the BLEU score and perplexity in the BART model. These results imply that direct-to-indirect transfer is more difficult than the indirect-to-direct task. As our statistical analyses presented in Section 3.3, indirect responses have a larger vocabulary and are longer on average. Hence, we conjecture that even the fine-tuned BART model does not acquire the ability to properly transform the direct response into an indirect one, regardless of the availability of the dialogue history. As seen from Table 9, the response generated by all models failed to preserve the intent of the direct response. More sophisticated models are desired to achieve direct-to-indirect transformation.

4.3 Directness Prediction Task

Task Description and Motivation This task aims to estimate the degree of directness of an utterance. This technology allows the rephrasing of utterances predicted as indirect into direct utterances using an indirect-to-direct transfer model or by asking users to clarify their intentions before inputting the utterance into a dialogue system.

In *DIRECT*, each dialogue history has a triple response: the original response from MultiWoZ, an indirect response, and a direct response. These responses can be ordered in descending order of directness as direct, original, and indirect responses. In this task, a model takes a response as an input and predicts the degree of directness.

Baselines We employ BERT as the baseline, where a response to predict its directness and dialogue history is input. The architecture is shown in Figure 7 (c). The output of the final layer corresponding to the '[CLS]' token is input into a linear layer, followed by a sigmoid function. The final output is regarded as the score indicating the directness of the response.

As discussed in Section 3.3, there is a remarkable difference in the frequency of words between the direct and indirect responses. We also employ a simple bag-of-words-based linear regression model to investigate the usefulness of word-level features in predicting directness.

Model	Loss	Exact	Kendall tau
BERT w/o history	Pointwise	0.785	0.803
BERT w/o history	Pairwise	**0.816**	**0.846**
BERT w/ history	Pointwise	0.784	0.804
BERT w/ history	Pairwise	0.813	0.841
LR w/o history	Pointwise	0.540	0.616

Table 10: Performance of the directness prediction task

We use pointwise and pairwise settings to train the model, which are typically used in learning-to-rank (Mitra and Craswell, 2018). Pointwise loss minimizes the mean squared error between the prediction and gold-standard directness scores. As the gold-standard, we set 1.0, 0.5, and 0.0 for the direct, original, and indirect responses, respectively.

Pairwise loss is designed such that the prediction score of a more direct response is larger than that of another. Suppose there is a direct response A and an indirect response B, whose predictions are s_A and s_B. The pairwise loss is defined as:

$$-\log \frac{1}{1 + e^{-(s_A - s_B)}}$$

As evaluation metrics, we compute the percentage of exact matches between a ranking based on the predicted scores and the gold standard. We also evaluate Kendall's tau between the prediction and gold standard and report the average.

We implemented the BERT-based model using 'bert-base-cased' with the transformers library, and the linear regression model using 'linear_model.LinearRegression' with the scikit-learn (version 0.23.1).[10] We also constructed a model that disregards the dialogue history for comparison.

Results and Analysis Table 10 shows the results. The BERT models disregarding dialogue history achieved higher scores than the models that use dialogue history. As revealed in Section 3.3, indirect and direct responses have largely different vocabulary and phrases, which may be clues for predicting the degree of directness. Nonetheless, the percentage of an exact match remains at 0.814 and a more sophisticated model is desired to effectively employ the dialogue history. The exact match rate for the linear regression model (LR w/o history) was 0.540, which was much higher than the chance rate of 0.167. This indicates that the bag-of-words-based model is useful for predicting directness, although it is not as good as BERT-based models.

Finally, we evaluated the performance of the response generation for the indirect and direct re-

[10] https://scikit-learn.org/stable/

sponses identified by the best model in Table 10. The model predicted 1,842 responses as indirect and 5,530 responses as direct.[11] We generated system responses to each of the indirect and direct user responses using the model proposed by Yang et al. (2021), which is the most advanced end-to-end response generation model for task-oriented dialogues. The BLEU scores of the indirect and direct responses were 10.25 and 14.09, respectively. This result confirms that direct utterances are easier for the dialogue system to respond accurately, while indirect utterances are more difficult.

5 Conclusion

We created *DIRECT*, a dialogue corpus providing 71,498 pairs of pragmatic paraphrases with context. In addition, we proposed three benchmark tasks and showed the performance of state-of-the-art pre-trained models as the baseline.

In a future, we will apply *DIRECT* to a task-oriented dialogue system to handle indirect responses in an end-to-end manner. We also intend to investigate the relations between pragmatic paraphrases and other features of dialogue acts and belief states.

Acknowledgements

This work was supported by the JSPS KAKENHI. Grant Number JP18K11435.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Penelope Brown, Stephen C. Levinson, and John J. Gumperz. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5016–5026.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pages 1–14.

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 (NAACL-HLT)*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP)*, pages 9–16.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 422–428.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 639–649.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing (PASCAL)*, pages 1–9.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1234–1245.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1224–1234.

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,

[11] Those with a score of 0.5 or higher were regarded as direct and the rest as indirect.

and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8.

Juliana Miehle, Wolfgang Minker, and Stefan Ultes. 2018. Exploring the impact of elaborateness and indirectness on user satisfaction in a spoken dialogue system. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP)*, page 165–172.

Bhaskar Mitra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.

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 (ACL)*, pages 311–318.

Louisa Pragst and Stefan Ultes. 2018. Changing the level of directness in dialogue using dialogue vector models and recurrent neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 11–19.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

John R. Searle. 1979. Indirect speech acts. In *Expression and Meaning: Studies in the Theory of Speech Acts*, page 30–57. Cambridge University Press.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 38–45.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards fully end-to-end task-oriented dialog systems with GPT-2. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 14230–14238.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 270–278.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1298–1308.

Yufan Zhao, Can Xu, and Wei Wu. 2020. Learning a simple and effective model for multi-turn response generation with auxiliary tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3472–3483.