Mimic and Rephrase: Reflective listening in open-ended dialogue

Justin Dieter
Square Inc. / Stanford University
justindieter@cs.stanford.edu

Tian Wang
Square Inc.
tian@squareup.com

Gabor Angeli
Square Inc.
gabor@squareup.com

Angel X. Chang*
Simon Fraser University
angelx@sfu.ca

Arun Tejasvi Chaganty*
Square Inc.
arun@squareup.com

Abstract

Reflective listening—demonstrating that you have heard your conversational partner—is key to effective communication. Expert human communicators often *mimic and rephrase* their conversational partner, e.g., when responding to sentimental stories or to questions they don't know the answer to. We introduce a new task and an associated dataset wherein dialogue agents similarly mimic and rephrase a user's request to communicate sympathy (*I'm sorry to hear that*) or lack of knowledge (*I do not know that*). We study what makes a rephrasal response good against a set of qualitative metrics. We then evaluate three models for generating responses: a syntax-aware rule-based system, a seq2seq LSTM neural models with attention (S2SA), and the same neural model augmented with a copy mechanism (S2SA+C). In a human evaluation, we find that S2SA+C and the rule-based system are comparable and approach human-generated response quality. In addition, experiences with a live deployment of S2SA+C in a customer support setting suggest that this generation task is a practical contribution to real world conversational agents.

1 Introduction

Humans in conversation naturally engage in reflective (or active) listening, where they indicate they have heard and understood their partner by repeating or rephrasing what they have said. This strategy has its roots in Rogerian psychology (Rogers, 1951) as a counseling technique meant to build trust and empathy. Dialog agents benefit from the same strategy to keep conversations pleasant, especially when the agent cannot help.

Reflective listening can be formalized into two aspects: (1) *mimicking and rephrasing* the conversational partner's utterance, and (2) incorporating an expressive speech act (Searle, 1976) appropriate for the utterance. For example, in Table 1 (b) we incorporate the speech act *I don't know* on top of the mimicked utterance *when the swimming pool is open*.

Prompt: *Hmm... I'm curious as to whether the swimming pool is open after 7pm?*
✓ a. I do not know that
✓ b. I don't know when the swimming pool is open.
✓ c. I don't know as to whether the swimming pool is open after 7pm.
✗ d. I don't know if you are curious as to whether the swimming pool is open after 7pm.
✗ e. I don't know Hmm... I'm curious as to whether the swimming pool is open after 7pm?

Table 1: Possible responses to indicate the dialog agent doesn't know the answer to a question (the prompt). The last two responses (d,e) are incorrect while the first three (a,b,c) are all acceptable with varying levels of specificity. The best response (b) is the one that is neither too vague (a) nor too verbose and repetitive (c).

In this paper, we propose a new task of generating mimic rephrasals for a given speech act. Table 1 illustrates the task with an example prompt and possible range of responses. The task is non-trivial to handle as naively putting the two parts together will result in responses that are either ungrammatical (e) or do not select the appropriate clause to rephrase (d). In our example, the first three responses all correctly convey the "I don't know" message. However, the blanket response (a) is overly vague and does not convey any understanding. Response (c) is specific but overly verbose and robotic. The best response is (b) since it contains enough details to signal understanding while remaining concise.

To get this best response, the agent must ig-

*These authors contributed equally.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 393–403
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

nore user flourishes (*"Hmm... I'm curious"*), identify the relevant portions of the prompt, correctly rephrase keywords (replace "I" with "you"), and coordinate arguments (using *"when"*).

By simulating reflective listening, we believe that mimic rephrasals will allow goal-oriented dialog systems to still respond naturally to open-ended input from users. In that vein, there has been renewed interest in open-ended dialog systems using neural models. However, Li et al. (2016a) have noted that naïve neural models tend to generate repetitive and dull responses such as "I don't know". While several attempts have been made to control various aspects of generation and hence produce more diverse output (Li et al., 2018; Hu et al., 2017; Logeswaran et al., 2018), we instead focus on expressing a single speech act (e.g., "I don't know"), but grounding it in diverse open-ended settings to simulate reflective listening.

In this paper, we examine what makes for a *good* response for a given speech act. We create two datasets IDONTKNOW and EMOTIVE focusing on two speech acts demonstrating reflective listening, respectively stating that we do not know an answer and expressing sympathy. We analyze the quality of responses along different dimensions such as fluency (is the response grammatical?), appropriateness (is the response on topic?), specificity, repetitiveness, and conciseness. We also compare responses from rule-based and neural models to gain insight into the strengths/weaknesses of different models at this task. We demonstrate that the rule-based model is repetitive but performs well for simple cases, while a sequence-to-sequence model with attention (Bahdanau et al., 2015) and a copying mechanism (Gu et al., 2016) has more varied responses and compares favorably. We release our dataset, code, and experiments to the community. [1]

2 Task: Mimic Rephrasals

In this section, we introduce the task of *Mimic Rephrasal* more formally. We use the term *speech act* to describe the information we want to convey to our conversational partner. For example, a lack of knowledge, sympathy, etc. We define a *prompt* as an utterance by a conversational partner that should trigger some form of speech act. For example, "where is my car?" could be a prompt for the speech act conveying a lack of knowledge. Note that detection of these prompts and classification into the appropriate speech act—while important for a real-world system—is outside the scope of this task.

The task is as follows: given a prompt and the target speech act, generate a *mimic rephrasal* of that prompt which conveys the speech act. We explore the two use-cases of rephrasing lack of knowledge (IDONTKNOW) and sympathy (EMOTIVE).

The important goal of the task to generate a response that makes the user feel that they have been heard and understood. Directly measuring this is difficult. Instead, we propose five metrics to characterize the quality of mimic rephrasals:

- **appropriateness** Did the response include the topic of interest in the rephrasal?
- **fluency** Is the response grammatical?
- **specificity** How much detail from the input prompt is captured?
- **conciseness** Is the response to the point?
- **repetitiveness** Is the response repetitive?

Looking at Table 1: (d) is not *appropriate* and (e) has low *fluency*. (a) to (c) are both *appropriate* and *fluent* with varying *specificity* (from low to high) and *conciseness* (from high to low). Intuitively, there is a tradeoff between specificity and repetitiveness: it should neither be too vague nor too repetitive.

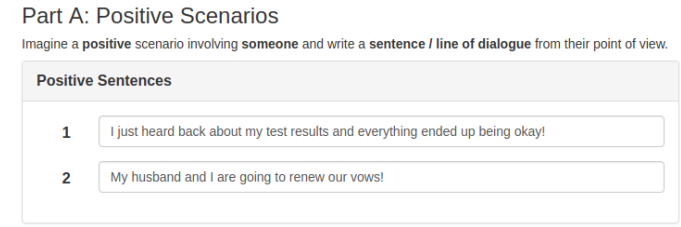
3 Dataset

Section 3.1 describes our process for collecting data, to document our dataset creation and to describe how to collect data for other speech acts—e.g., expressing gratitude, soliciting confirmation of intents, etc. The subsequent section (Section 3.2) describes some statistics of the two datasets in this paper: EMOTIVE and IDONTKNOW.

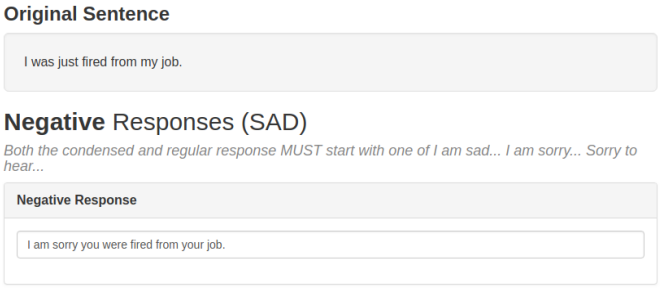
3.1 Data collection

Our data collection pipeline has two phases. In the first phase, workers were asked to come up with a *prompt* or scenario. For example, a question to ask the dialog agent, or a sentiment laden scenario. During this phase, workers were asked to be as creative as possible, to explore a variety of sentence structures and lengths in the way they phrase their prompts, and diversity in the topics covered. In the second phase, we asked another set of workers to generate multiple *responses* each to the prompt.

[1] https://github.com/square/MimicAndRephrase/



(a) Interface when generating prompts



(b) Interface when generating mimic rephrasals

Figure 1: Our data collection pipeline. First, crowdworkers are asked to generate a given *prompt* or scenario (a). Then, a different worker is asked to generate the *mimic rephrasal* of the prompt generated by the first worker (b).

IDontKnow
Prompt: I am legally resident in Northern Ireland, where can I apply for an Irish visa **Mimic Rephrasal**: I do not know where you can apply for an Irish visa
P: When I register a domain, do I receive a website and a web hosting space **MR**: I do not know if you receive a website and a web hosting space when you register a domain
P: I'm having difficulty signing up. Whom can I contact **MR**: I do not know who you can contact about your difficulty signing up
Emotive
P: The sisters were able to reunite after 20 years **MR**: I am happy to hear the sisters were able to reunite after all this time
P: The future looks brighter than I ever imagined **MR**: I 'm happy that your future looks bright to you
P: My phone fell into the toilet and it's ruined now. **MR**: I am sad that your phone is ruined because it fell into the toilet

Table 2: Examples of mimic rephrasals in the IDontKnow and Emotive datasets collected in this paper. Each example has a *prompt* (the utterance from the conversational partner), and a *mimic rephrasal*: the utterance that should be returned by the dialog agent.

We found that splitting the task up into these two steps—generating prompts and then responses—improved the quality of our collected sentence pairs.

The interface used by Mechanical Turk workers is shown in Figure 1. Workers are first asked to generate a number of prompts (scenarios) in Figure 1 (a). Once these prompts are collected, a different set of workers were asked to generate responses to the prompts, completing our dataset (see Figure 1 (b)). Workers were paid \$0.10 per sentence in the prompt generation task, and \$0.07 per sentence in the response generation task.

3.2 Dataset statistics

We use the method described in Section 3.1 to collect two datasets IDontKnow and Emotive. Both datasets are split into train/dev/test splits with a ratio of 70/15/15%.

IDontKnow is a dataset for indicating that we don't know the answer to a question, or cannot execute a request. **Emotive** is a dataset for expressing sympathy for the topic of the prompt, with a balanced distribution of positive and negative sentiment. Examples for the two datasets can be found in Table 2, and statistics are given in Table 3. The modest size of the training set (10 189) means that a good fraction of the test set contains out of vocabulary words: the 1 377 test examples contain 512 words not seen during training, motivating our use of a copy mechanism.

We report statistics both on the full mimic rephrasal (MR), as well as for just the *Mimic* por-

	IDK	EMOTIVE
Dataset size		
Training pairs	6435	3887
Development pairs	1377	834
Test pairs	1377	828
Sentence length		
Prompt mean token count	11.7	11.0
Mimic mean token count	8.8	9.1
MR mean token count	12.9	10.2
Train Vocabulary		
Prompt vocabulary size	5703	4060
MR vocabulary size	5136	3200
Prompt/MR Jaccard Sim	0.83	0.61

Table 3: Statistics about the IDONTKNOW and EMOTIVE datasets. *Mimic* is here taken to be the mimicked utterance, without the preceding "I am happy/sorry/etc." or "I don't know". *MR* is the complete mimic rephrased response.

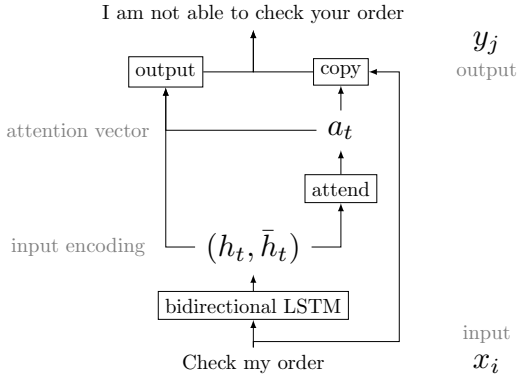


Figure 2: An outline of the sequence to sequence model with attention and a copying mechanism. The input phrase is *"Check my order"* with a correct output of *"I am not able to check your order"*.

tion of the mimic rephrasal. For example, for the MR "I do not know where you can apply for an Irish visa", the *Mimic* portion would be "where you can apply for an Irish visa." While the average MR is slightly longer than the original prompt, the Mimic portion averages 2.9 tokens (25%) shorter than the prompt. In addition, the high Jaccard similarity between the prompt and mimic portion suggests the task involves selecting key portions of the original sentence. [2]

4 Methods

In this section we describe three models for the task described above. These include a rule based baseline constructed with deterministic syntactic transformations as well as trained neural models.

4.1 Rule based baseline

As a naïve baseline, we use a set of hand-written syntactic rephrasing rules using Stanford CoreNLP (Manning et al., 2014). For example, for the IDONTKNOW dataset we developed 8 rules that match a Semgrex (Chambers et al., 2007) pattern to an associated dependency graph manipulation algorithm. For example, the rule based system matches the phrase *"What is the difference between the debt and the deficit?"* to a general type of pattern where the verb (in this case *"is"*) needs to be extracted from the dependency graph and reattached at the end to produce *"I do not know what the difference between the debt and the deficit is."* The EMOTIVE rule based system extracts the root clause from the constituency parse of the sentence and adds enclosing phrasing ("Sorry to hear that..."). Both systems also use simple string manipulation to replace pronouns and correct casing. We note that the rules were developed by iterating on the training data. In the appendix, we include a histogram of how often each rule was fired in the IDONTKNOW rule based system.

Although this is a strong baseline, it has some weaknesses. Writing and maintaining the rule set is difficult and time consuming. The Semgrex patterns and accompanying transformations are non-trivial and requires expert time to develop and maintain. Additionally, it is difficult to deterministically decide which portions of the prompt to keep or drop in the rephrasal. For instance, *"I found out someone has been stealing from me"* should drop the *found out* and respond with: *"Sorry to hear that someone has been stealing from you"*.

4.2 Neural Models

To address these issues, we develop a series of neural models for the task. Formally, let $\mathbf{x} = x_1, \ldots, x_n$ be the source sentence, and $\mathbf{y} = y_1, \ldots, y_m$ be the generated sequence of output tokens. We define a seq2seq model similar to existing neural MT models (Cho et al., 2014) for generating $\mathbf{y}$ given an input $\mathbf{x}$, as well as a model augmented with a copy mechanism.

Input embedding. All models use concatenated ELMo (Peters et al., 2018) and GloVE (Pennington et al., 2014) embedding for the input embeddings. The model architectures used for the EMOTIVE and IDONTKNOW datasets are identical with one

[2] Jaccard similarity is computed as the intersection over union of lemmatized non-stopword tokens between the prompt and Mimic portion.

exception. For the EMOTIVE task, an extra bit is appended to the word embeddings to specify whether the input has a positive or negative sentiment. [3]

Baseline neural model. Our baseline neural model is a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with attention (Bahdanau et al., 2015). Formally, the *encoder* outputs a set of hidden states for each token given by $\{\bar{h}_1, \ldots, \bar{h}_n\} = LSTM\{m(x_1, \ldots, x_n)\}$ for word embeddings $m(x)$. The decoder is also an LSTM with hidden state initialized to the sum of the final hidden states of the forward and backward LSTMs contained in the bidirectional LSTM encoder. For encoder hidden state $\bar{h}_s$ and decoder hidden state h_t, the attention score $a_t(s)$ is defined as

$$a_t(s) = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_{s'}))}$$

where $\text{score}(h_t, \bar{h}_s) = \boldsymbol{v}_a^T \cdot \tanh(W_a[h_t; \bar{h}_s])$ and $\boldsymbol{v}_a$ and W_a are learnable parameters.

Decoding uses a modified beam search (see Section 4.3),[4] and the model is trained on the following cross entropy loss function:

$$J = \sum_{i=1}^{m} -\log p_v(y_i = y_i^* | y_{<i}, x)$$

where y_i^* is the expected output token given in the training data and $\log p_v(y_i = y_i^* | y_{<i}, x)$ is a distribution over the models vocabulary computed from the logits outputted by the model.

Neural model with copying. The most effective neural model we implemented augments the baseline neural model with a copying mechanism (Gu et al., 2016). This allows the model to generalize better to unseen vocabulary, and more strongly enforces the core tenant of the task: that we should be *mimicking* the prompt. An overview of the model is shown in Figure 2.

The key difference from the baseline neural model is that we now generate output tokens using a combined softmax over the model's vocabulary and the tokens in the input:

$$\log p(y_i = y_i^* | y_{<i}, x) = \log p_v(y_i = y_i^* | y_{<i}, x) + \sum_{\{j | x_j = y_i^*\}} \log p_c(y_i = \text{copy}(x_j) | y_{<i}, x),$$

where $\log p_v(\cdot)$ is the same as before, and $\log p_c(\cdot)$ is a distribution over the words in the input. For further details we defer the reader to (Gu et al., 2016). Similar to the baseline model, we decode the model using a modified beam search and train it using a cross-entropy loss.

4.3 Modified Beam Search

We use a modified version of beam search (Huang et al., 2017) when generating output tokens to favor longer responses. The modified beam search first calculates the average ratio of output tokens to input tokens from the dev set, k. We then compute the average logit value of an individual output token, $r(y_i)$, over all outputs produced on the dev set input, r_{avg}. A modified perplexity, $\tilde{sc}(\mathbf{y}, \mathbf{x})$, is used to determine which beams to prune, where $\mathbf{x}$ is a series of input tokens and $\mathbf{y}$ is a proposed series of output tokens (a beam):

$$\tilde{sc}(\mathbf{y}, \mathbf{x}) = sc(\mathbf{y}) + r_{avg} \cdot \min\{\text{len}(\mathbf{y}), k \cdot \text{len}(\mathbf{x})\}$$

where $sc(\mathbf{y}) = \sum_{i=1}^{\text{len}(\mathbf{y})} r(y_i)$ is the standard perplexity for the generated output.

Additionally, we found that for a given output $\bar{\mathbf{y}}$, the score $\tilde{sc}(\bar{\mathbf{y}}, \mathbf{x})$ provides a good measure of generation quality and is useful when filtering out poor or unacceptable output.

4.4 Training

All models were implemented and trained using PyTorch (Paszke et al., 2017). The Adam (Kingma and Ba, 2014) optimizer was used for all gradient based optimization.

We used a randomized hyperparameter grid search to determine the learning rate, number of layers, dropout, and the dimensions of the hidden layers. We used a learning rate of 0.000718 for all optimization. A dropout value of 0.1 is used for all models. All LSTMs are bidirectional with a single layer. Both sequence to sequence models for the IDONTKNOW task use a hidden size of 524 within the LSTM, a hidden size of 100 for the attention layer, a hidden size of 638 for the copy layer, and a dropout value of 0.1. Both sequence to sequence models for the EMOTIVE task use a hidden size of 600 within the LSTM, a hidden size of 200 for the attention layer, a hidden size of 650 for the copy layer, and a dropout value of 0.1.

5 Experiments

We study what makes a good response by correlating human judgments of goodness (based on a

[3] We note that the rule-based system simply used different rules for different settings of this bit.

[4] Initial experiments show that this modified beam search worked better

Model	IDontKnow Dev	IDontKnow Test	Emotive Dev	Emotive Test
BLEU				
Rule-based	78.9	79.4	**47.0**	**46.6**
S2SA	63.3	63.1	32.9	34.2
S2SA+C	**79.9**	**79.7**	44.6	46.3
METEOR				
Rule-based	84.6	85.3	**54.4**	**54.1**
S2SA	74.1	73.6	37.1	38.1
S2SA+C	**88.4**	**88.5**	51.5	52.8

Table 4: BLEU and METEOR scores evaluated on the Dev and Test sets. The S2SA+C performs the best on IDontKnow and the rule-based performs the best on Emotive.

5-point Likert scale) to the set of qualitative metrics we defined in Section 2. The average score is 4.2 for IDontKnow and 3.7 for Emotive. We also evaluated the performance of different models on these datasets using: (1) an automated BLEU and METEOR evaluation, (2) an A/B study comparing the model output with the gold response, and (3) the qualitative metrics. We conclude with a qualitative error analysis and some observations from a live deployment of the neural rephrasal model.

We compare the responses generated by three models: (1) the **Rule-based** baseline; (2) **S2SA**: neural model consisting of a seq2seq model with attention, and (3) **S2SA+C**: neural seq2seq model with attention and copying.

5.1 Results

BLEU/METEOR Following prior work on text generation, we use BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) to compare the performance of our model.[5] From the results shown in Table 4, the neural model with copying (**S2SA+C**) are the rule-based baseline have comparable performance, with both significantly outperforming the baseline neural model (**S2SA**).

A/B Test We perform a human evaluation of our model outputs by creating an A/B test where evaluators specify a preference for either the model output, or the gold human response (see Figure 3). A perfect score on this evaluation would be 50%, indicating that the model and human response are indistinguishable. We ran this test on 305 examples selected from the test set. For each example which was not identical to the gold output, five Turk Workers were asked to choose which response they preferred. Table 5 shows the result of the study, showing that the copy mechanism outperforms both the rule-based baseline and the **S2SA** model for IDontKnow. For Emotive, the rule-based baseline is preferred.

[5]Specifically, we used NLTK (Loper and Bird, 2002) to compute both BLEU and METEOR scores with one human reference for each example. The BLEU score is weighted equally between 1-grams, 2-grams, 3-grams, and 4-grams

Model	IDontKnow P%	IDontKnow (I%)	Emotive P%	Emotive (I%)
Rule-based	38.0	(28.9)	**39.7**	(0.25)
S2SA	16.4	(11.5)	12.1	(0.25)
S2SA+C	**46.7**	(44.3)	35.9	(1.0)

Table 5: The percent of model responses that were (P)referred over the human responses in the A/B test portion of the user study on 305 IDontKnow examples and 400 Emotive examples. When the model's response was identical to the human's, we assume it is preferred 50% of the time: the percentage of these examples is reported in the (I) column.

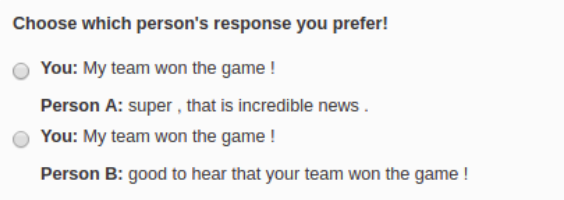


Figure 3: Example question in A/B test. The prompt asks the crowdworker to choose between the human response and the model output. In this example, one would prefer the Person B's response because it is more specific and exhibits reflective listening.

Qualitative Metrics To gain insight into the types of errors the different models are making, we elicited human assessment of three of the metrics defined in Section 2:

1. **Appropriateness** Evaluators make a binary choice as to whether the response included the correct part of the prompt.
2. **Fluency** Evaluators assess the grammatical correctness of the response by selecting on a 3 point Likert scale ranging from "not fluent" to "somewhat fluent" to "fluent". Scores are normalized to 1.
3. **Specificity** We present evaluators with a response and ask them to pick the original prompt from 4 choices (the original prompt, two distractors, and "none/multiple applies"). The distractors are chosen from the nearest neighbors of the prompt using an averaged

Model	App.	Flu.	Spec.	Con.	Rep.
IDONTKNOW					
Human	*93.1*	*91.75*	*74.7*	*1.18*	*65.1*
Rule-based	86.3	85.2	79.3	1.25	74.4
S2SA	49.5	66.8	24.4	1.12	45.7
S2SA+C	**92.4**	**86.0**	**77.5**	**1.22**	**70.4**
EMOTIVE					
Human	*93.2*	*88.5*	*61.3*	*0.97*	*36.3*
Rule-based	**94.4**	**91.7**	90.0	1.29	84.4
S2SA	29.8	74.4	7.0	**0.97**	17.7
S2SA+C	76.4	76.1	**63.5**	1.04	**49.7**

Table 6: Assessment and measures of **App**ropriateness, **Flu**ency, **Spec**ificity, **Con**ciseness and **Rep**etitiveness. We **bold** the highest scores for App. and Flu., and closest to human for Spec., Con., and Rep.

Model	App.	Flu.	Spec.	Con.	Rep.
IDONTKNOW					
Human*	**0.34**	0.14	0.17	**-0.39**	**0.38**
Rules	**0.39**	**0.42**	0.17	-0.03	0.08
S2SA	**0.54**	**0.30**	**0.35**	0.00	0.23
S2SA+C	**0.45**	**0.48**	0.09	-0.19	0.21
EMOTIVE					
Human*	**0.52**	**0.59**	0.05	-0.17	-0.08
Rules	0.12	0.04	-0.08	-0.08	0.01
S2SA	**0.63**	0.23	0.22	-0.03	**0.33**
S2SA+C	**0.41**	**0.41**	0.00	0.02	-0.00

Table 7: Correlations between diagnostic metrics and human quality judgments for responses in the two datasets, with **bold** indicating statistically significant correlations. For all model responses, we use the A/B test preferences as a measure of quality judgment. For the human responses, we use a 5-pt Likert scale of the "goodness" of response as a quality judgment.

GloVE sentence embedding.

We also used the following automatic metrics as proxies for the remaining qualitative metrics:

1. **Conciseness** The length of the response normalized to the length of the prompt (smaller is more concise).
2. **Repetitiveness** We use METEOR (Banerjee and Lavie, 2005) to measure the overlap between prompt and response.

Table 6 shows the results of the human judgment on 400 generated responses for the test set. The **rule-based** model does well on the EMOTIVE dataset. The **S2SA** model is overall worst on most metrics, except conciseness. On the other hand, the **S2SA+C** model performs best on appropriateness and fluency for the IDONTKNOW dataset, and compares favorably with the rule-based model for other metrics. We note that the **S2SA+C** model most closely matches the amount of specificity, conciseness, and repetitiveness in the human responses. Examples of human responses with their corresponding metrics is provided in the appendix, along with additional responses from the models and error analysis.

5.2 Analysis

Next, we look at the correlation of our qualitative metrics to overall human quality judgments. Results of this analysis are in Table 7 and we provide additional visualizations in the appendix. The human response goodness score correlated positively with appropriateness, fluency, and to some extent with repetitiveness. On the other hand it correlated negatively with conciseness (i.e., shorter responses are preferred), while correlation with specificity was less pronounced. This seems to indicate that good responses are characterized by being appropriate and fluent, while having an appropriate level of detail (indicated by some amount of repetitiveness balanced with conciseness). We see a similar trend for the model responses: appropriateness and fluency are the most important attributes for when a model's response is preferred over the human's.

To get a better qualitative understanding of the model's performance, we studied the responses generated by our models (see Table 8 for examples). For simple sentences, the IDONTKNOW **rule-based** responses are reasonable. However, for more complex sentences, it becomes challenging to identify relevant subclauses or to handle non-trivial constructions like conditional clauses (see aquarium example). As a result, the EMOTIVE **rule-based** responses, while grammatical, tend to be overly verbose.

Both **S2SA** and **S2SA+C** are good at producing relatively fluent sentences, and performing the correct pronoun replacements ("you" with "we" and "I"). **S2SA** responses are often off-topic and inappropriate, with the model generating words that related to the topic but prone to drift (e.g., the teapot example). Since we train on a very small dataset, many words in the prompt are not seen during training. While the input word embedding can help during encoding, the decoder is nonetheless unable to generate words it has not seen during training.

The two biggest errors **S2SA+C** makes are incorrectly identifying the relevant parts of the question

	IDONTKNOW	EMOTIVE
Prompt	If I win my case, what am I entitled to	I dropped the jar and it shattered everywhere
Human	I do not know what you are entitled to	I am sorry to hear the jar broke
Rule-based	I do not know what you are entitled to	I am sorry you dropped the jar and it shattered everywhere
S2SA	I do not know what you are required to	i am sorry to hear about the ice
S2SA+C	I do not know what you are entitled to	i am sorry you dropped the jar
Prompt	At what temperature should I heat my water in a Staub teapot	The dentist told me that my insurance did n't cover dental
Human	I do not know what temperature you should heat your water to in a Staub teapot	I am sorry your insurance does n't cover dental
Rule-based	I do not know at what temperature you should heat your water in a Staub teapot	I am sad the dentist told you that your insurance did n't cover dental
S2SA	i am not able to tell you what long it you should handle your tea in a tsunami machine	sorry to hear that you ca n't afford your insurance insurance insurance insurance
S2SA+C	i do not know at what you should heat your water in a staub teapot	sorry to hear that the dentist did n't cover dental
Prompt	Are the animals at your aquarium humanely treated	Did I tell you that I won $ 500 at bingo
Human	I do not know if the animals at our aquarium are humanely treated	I am glad you won it
Rule-based	I do not know of the are animals at my aquarium humanely treated	I'm happy did you tell me that you won $ 500 at bingo
S2SA	i do not know if the animals are at our zoo are allowed	i am happy you won the lottery
S2SA+C	i do not know if the animals are at our aquarium are treated treated	i am happy that you won $ 500 at bingo

Table 8: Example responses from the different models, with errors highlighted in red. Note that the **S2SA** model tend to introduce random terms and **S2SA+C** model will retain numbers such as $ 500.

(e.g., the response *"I do not know what reporter's transcript deposit are"* to the question *"What are Reporter's Transcript Deposit Costs?"*) and grammatical errors when rephrasing (e.g., the response *"I do not know when the free trial is end."* to the question *"When does the free trial end?"*).

5.3 Observations from a Live Deployment

We deployed the **S2SA+C** model as part of a live chatbot for customer service that helped answer customer queries and perform simple tasks like tracking their packages.[6] As context, customers would ask the chatbot questions (e.g., "how do I ship pets?" or "I want to change the delivery address for my package") which were matched against a knowledge base containing frequently asked questions. If a question similarity model was unable to find a match, we tried to communicate to the user that we could not answer their request. Prior to this work, the chatbot would respond with a generic backoff message: "I'm sorry I didn't understand something you said" which resulted in users repeatedly rewording their request even if their request was genuinely outside of the chatbot's knowledge base, and ultimately expressing frustration with the chatbot.

We incorporated mimic rephrasals into our system by responding with the output generated by the S2SA+C model trained on the IDONTKNOW dataset if its score, $\tilde{sc}(\mathbf{y}, \mathbf{x})$, was higher than a fixed threshold, and using the previous backoff response if not. We observed that when the model replied with a mimic rephrasal, users usually responded with gratitude, e.g. "Thanks for letting me know!", and either continued by asking a different question or leaving the conversation. Presented with the mimic response, users rarely wasted time rewording a request that was out of the scope of what the chatbot could handle.

6 Related Work

Verbal mimicry is used in conversation to build social rapport (Rogers, 1951; Rautalinko and Lisper, 2004). This observation has been leveraged even in the early development of conversational agents, for example in systems such as Eliza (Weizenbaum, 1966), which engaged users by picking up keywords and repeating open ended questions back to the patient, or PARRY (Colby et al., 1972), which follows a similar strategy to rephrase utterances to

[6]The live deployment was run at Eloquent Labs prior to their acquisition by Square Inc.

indicate anger or fear. Our work applies mimicry in the task-oriented setting and studies how and when generated mimic rephrasals are preferred over human responses.

More recently, sequence to sequence models have been used for open domain chatbots (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015; Wen et al., 2015; Serban et al., 2016). However, these models suffered from generic responses and turns that are semantically inconsistent and incoherent. To address these issues, Li et al. (2016a) introduced a maximum mutual information objective to encourage diversity. Consistency in dialog agents has been well studied in Kobsa and Wahlster (1989) *inter-alia*, and for neural methods by Li et al. (2016b). Rashkin et al. (2019) also recognized the need to acknowledge others' feelings in a conversation and introduced a dataset for benchmarking emphathetic dialog models.

Sequence-to-sequence models with copying was introduced in Gu et al. (2016). Such models have also been shown to be effective at semantic parsing (Jia and Liang, 2016), summarization (See et al., 2017; Cao et al., 2018), and task oriented dialog (Eric and Manning, 2017).

Our task can in many ways be considered a controlled generation task. Other work in this area includes generating text conditioned on a sentiment to express (Li et al., 2018), or controlled generation (Hu et al., 2017) by editing attributes (Shen et al., 2017; Logeswaran et al., 2018; Lample et al., 2019). These works can successfully change the tone and intent of an utterance, but tend to frequently rewrite enough of the content that the method is less effective for practical dialog applications. Ke et al. (2018) examined how to generate dialog responses with different sentence function (e.g., imperative, interrogative, etc.), which similarly allows for more distant rewriting than is optimal for our task. Finally, our task exhibits many of the challenges observed by Bilu et al. (2015) in the context of negating claims.

Other work in generation addresses tasks such as rephrasals for generating paraphrases (Prakash et al., 2016; Gupta et al., 2018), sentence simplification (Narayan et al., 2017), and query rewriting for question answering (Dong et al., 2017).

7 Conclusions

We proposed a new task and associated datasets for *mimicking and rephrasing* a speaker's prompt to communicate a given intent. We showed that both rule-based based and neural seq2seq models both approach human level performance. Additionally, we share observations from a real world deployment of the model to highlight how solving these tasks can potentially improve the end-user experience. We hope this will inspire future work in dialog agents, making these agents more fluent and personable.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. 2015. Automatic claim negation: Why, how and when. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine De Marneffe, Daniel Ramage, Eric Yeh, and Christopher D Manning. 2007. Learning alignments and leveraging natural logic. In *ACL-PASCAL Workshop*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kenneth Mark Colby, Franklin Dennis Hilf, Sylvia Weber, and Helena C. Kraemer. 1972. Turing-like indistinguishability tests for the calidation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3:199–221.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mihail Eric and Christopher Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the Association for the Advancement on Artificial Intelligence (AAAI)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Liang Huang, Kai Zhao, and Mingbo Ma. 2017. When to finish? optimal beam search for neural text generation (modulo beam size). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating informative responses with controlled sentence function. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alfred Kobsa and Wolfgang Wahlster. 1989. *User models in dialog systems*. Springer.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. *North American Association for Computational Linguistics (NAACL)*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Shashi Narayan, Claire Gardent, Shay B Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*.

Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and

dataset. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Erik Rautalinko and Hans-Olof Lisper. 2004. Effects of training reflective listening in a corporate setting. *Journal of Business and Psychology*, 18(3):281–299.

Carl R. Rogers. 1951. *Client-centered therapy*. Riverside Press.

John R Searle. 1976. A classification of illocutionary acts. *Language in society*, 5(1):1–23.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Association for the Advancement on Artificial Intelligence (AAAI)*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of ACL-IJCNLP*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the International Conference on Machine Learning, Deep Learning Workshop*.

Joseph Weizenbaum. 1966. ELIZA —a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.