Human-like informative conversations: Better acknowledgements using conditional mutual information

Ashwin Paranjape
Stanford University
ashwinp@cs.stanford.edu

Christopher D. Manning
Stanford University
manning@cs.stanford.edu

Abstract

This work aims to build a dialogue agent that can weave new factual content into conversations as naturally as humans. We draw insights from linguistic principles of conversational analysis and annotate human-human conversations from the Switchboard Dialog Act Corpus to examine humans strategies for *acknowledgement*, *transition*, *detail selection* and *presentation*. When current chatbots (explicitly provided with new factual content) introduce facts into a conversation, their generated responses do not ***acknowledge*** the prior turns. This is because models trained with two contexts – new factual content and conversational history – generate responses that are non-specific w.r.t. one of the contexts, typically the conversational history. We show that specificity w.r.t. conversational history is better captured by *pointwise* ***conditional*** *mutual information* (pcmi_h) than by the established use of *pointwise mutual information* (pmi). Our proposed method, Fused-PCMI, trades off pmi for pcmi_h and is preferred by humans for overall quality over the Max-PMI baseline 60% of the time. Human evaluators also judge responses with higher pcmi_h better at acknowledgement 74% of the time. The results demonstrate that systems mimicking human conversational traits (in this case acknowledgement) improve overall quality and more broadly illustrate the utility of linguistic principles in improving dialogue agents.

1 Introduction

Social chatbots are improving in appeal and are being deployed widely to converse with humans (Gabriel et al., 2020). Advances in neural generation (Adiwardana et al., 2020; Roller et al., 2020) enable them to handle a wide variety of user turns and to provide fluent bot responses. People expect their interactions with these dialogue agents to be similar to real social relationships (Reeves and Nass, 1996). In particular, they expect social chatbots to both use information that is already known and separately add new information to the conversation, in line with the *given-new* contract (Clark and Haviland, 1977).

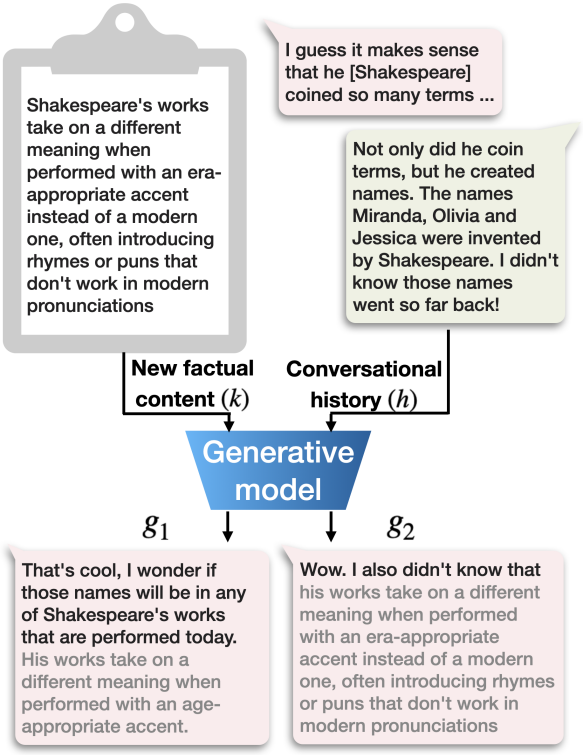

Figure 1: **The setting for conversational rephrasing.** Conversational history (**h**) and new factual content (**k**), two largely independent contexts, are used to sample responses ($\mathbf{g}_1$, $\mathbf{g}_2$) from a generative model. The samples differ qualitatively. While almost all of $\mathbf{g}_2$ is verbatim from **k** (in gray), the first sentence in $\mathbf{g}_1$ (in black) acknowledges using **h** and bridges to **k**.

Neural generation methods for adding new information (Dinan et al., 2019; Gopalakrishnan et al., 2019; Ghazvininejad et al., 2018; Zhang et al., 2018) measure progress using metrics like "engagingness", "appropriateness" and "informativeness". But these metrics are too broad and provide little actionable insight to drive improvements in these systems. On the other hand, psycholinguists and sociolinguists have studied human conversations in

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 768–781
June 6–11, 2021.

depth and have identified fine-grained conventions, principles and contracts (Grice, 1975; Clark, 1996; Krauss and Fussell, 1996).

Our first contribution is a linguistic analysis of how human conversations incorporate world knowledge. We manually annotate conversations from the Switchboard corpus to identify key traits. In particular, we find that people apply four kinds of strategies: (1) **acknowledgement** of each other's utterances, (2) **transition** to new information, (3) appropriate level of **detail selection** and (4) **presentation** of factual content in forms such as opinions or experiences.

To identify deficiencies of the above types in machine-learned models, we consider a simplified task of **conversational rephrasing** (Figure 1), in which the factual content to be added is not left latent but is provided as a text input to the model (as in Dinan et al. (2019)), along with conversational history. Just as humans do not recite a fact verbatim in a conversation, we expect the model to rephrase the factual content by taking conversational context into account. We derive the data for this task using the Topical Chat dataset (Gopalakrishnan et al., 2019) and fine-tune a large pre-trained language model on it.

Li et al. (2016); Zhang et al. (2020) use maximum pointwise mutual information (Max-PMI) to filter out bad and unspecific responses sampled from a generative language model. However, we observe that Max-PMI responses lack in acknowledgement, an essential human trait. This is because a generated response that simply copies over the new factual content while largely ignoring the conversational history can have high mutual information (MI) with the overall input.

Our second contribution is a method to select responses that exhibit human-like acknowledgement. To quantify the amount of information drawn from the two contexts of new factual content and conversational history, we propose using **pointwise conditional mutual information (PCMI)**. We show that responses with a higher PCMI w.r.t conversational history given factual content (pcmi_h) are judged by humans to be better at acknowledging prior turns 74% of the time.[1] Then, we use pcmi_h to identify Max-PMI responses that lack acknowledgement and find alternative responses (Fused-PCMI) that trade off pmi for pcmi_h. Despite a lower PMI, human annotators prefer the Fused-PCMI alternative over the Max-PMI response 60% of the time.[1] We release annotated conversations from the Switchboard corpus (with guidelines), code for fine-tuning and calculating scores and human evaluations.[2]

Figure 2: Examples for **Acknowledgement Strategies** from Switchboard (parts omitted for brevity).

2 Strategies for informative conversations

To understand strategies used by humans while talking about factual knowledge, we annotate turns in human-human conversations. We adopt and extend Herbert Clark's approach to conversational analysis. According to his *given-new* contract (Clark and Haviland, 1977), the speaker connects their utterances with the given information (assumed to be known to the listener) and adds new information. This builds up *common ground* (Stalnaker, 2002) between the two participants, defined to be the sum of their mutual, common or joint knowledge, beliefs and suppositions. We identify the following four aspects to the process of adding new information to a conversation.

Acknowledgement strategies According to Clark and Brennan (1991), the listener provides positive evidence for grounding. We classify all mentions of prior context into various acknowledgement strategies.

Transition strategies According to Sacks and Jefferson (1995), topical changes happen step

[1] Statistically significant with $p < 0.05$ (Binomial Test).

[2] https://github.com/AshwinParanjape/human-like-informative-conversations

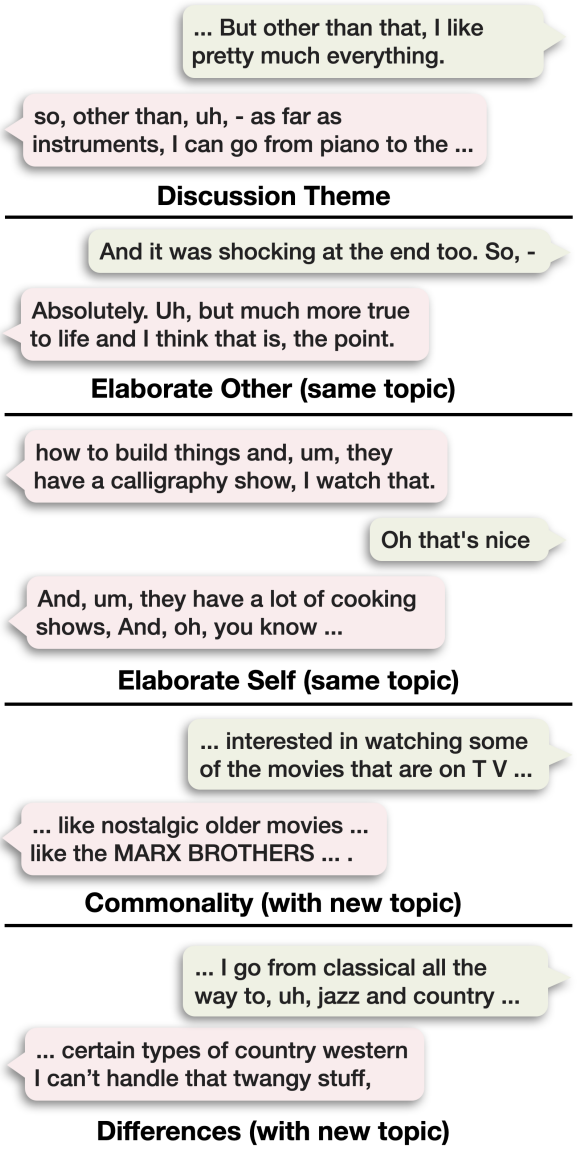

Figure 3: Examples for popular **Transition Strategies** from Switchboard (parts omitted for brevity).

by step, connecting the given, stated information to new information. We annotate the semantic justifications for topical changes as different transition strategies.

Detail selection strategies According to Isaacs and Clark (1987), speakers in a conversation inevitably know varying amounts of information about the discussion topic and must assess each other's expertise to accommodate their differences. We posit that each speaker applies detail selection strategies to select the right level of detail to be presented.

Presentation strategies According to Smith and Clark (1993), presentation of responses is guided by two social goals – exchange of information and self-presentation. While we do not consider social goals in this work, we hypothesize that people talk about factual information in non-factual forms (e.g., opinions, experiences, recommendations) which we classify as various presentation strategies.

2.1 Analysis of strategies

Dataset We annotate part of the The Switchboard Dialog Act Corpus (Stolcke et al., 2000), an extension of the Switchboard Telephone Speech Corpus (Godfrey et al., 1992) with turn-level dialog-act tags. The corpus was created by pairing speakers across the US over telephone and introducing a topic for discussion. This dataset is uniquely useful because as a speech dataset, it is more intimate and realistic than text-based conversations between strangers. We annotate conversations on social topics which might include specific knowledge (like Books, Vacations, etc.) but leave out ones about subjective or personal experiences.

Specific knowledge We define *specific knowledge* as knowledge that can be "looked up" but isn't widely known (as opposed to *general knowledge* that everybody is expected to know and *experiential knowledge* that can only be derived from embodied experiences). In this work, we are interested only in specific knowledge because it serves as a source of new information in a conversation that is hard for a language model to learn implicitly but is likely available as text that can be supplied to the system. Out of 408 annotated turns, 111 (27%) incorporate specific knowledge and account for 56% of the tokens.

Next, we analyze various strategies employed in turns containing specific knowledge:

Acknowledgement Strategies In 70% of the turns, the speaker acknowledges the prior turn corroborating Clark and Brennan (1991). Three main strategies (Figure 2): ***agreement*** (or disagreement), ***shared experiences*** (or differing experience) and ***backchanneling*** account for 60% of the turns (Figure 4). In certain cases, explicit acknowledgement isn't necessary. For example, the answer to a question demonstrates grounding and serves as an implicit acknowledgement. These are categorized as ***N/A***.

Transition Strategies At the beginning of a conversation, the participants use the ***discussion theme*** to pick a topic (various transition strategies are shown in Figure 3). The decision to stay on the topic or to transition to a new one is an implicit form of negotiation and depends on the interest

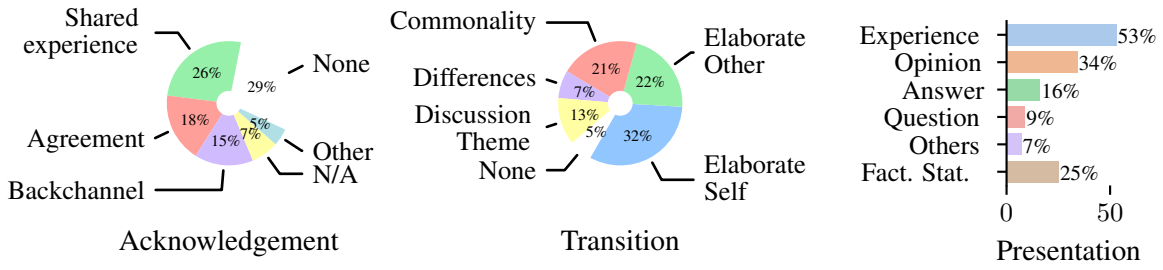

Figure 4: Distribution of acknowledgement, transition and presentation strategies

and ability of both speakers to participate. Nearly half the time, people elaborate upon the current topic (Figure 4). With a supportive listener, they might elaborate upon their own prior utterance (***self-elaboration***). Or they might signal interest in continuing the topic by elaborating the other speaker's utterance (***other-elaboration***). However, in a quarter of the turns, a participant loses interest or both participants run out of material. In that case, they transition to a new topic, implicitly justified by ***commonalities*** or ***differences*** with the current topic. If all else fails, they fall back to the ***discussion theme*** to pick a new topic.

Detail-selection strategies People probe the other speaker's knowledge about an entity before diving into details. As a probing mechanism, people introduce an entity without any details (***introduce-entity***) 50% of the time. Depending on the response, ***details*** are laid out 66% of the time. Note that a turn can have both labels, i.e., it can introduce an entity for the first time or it can have details of one entity while also introducing another entity. Interestingly, in 7% of turns, an entity's name is omitted but some details are presented, creating an opening for the other speaker to chime in.

Presentation strategies A single utterance can have multiple modes of presentation. A ***factual*** (objective) statement of specific knowledge is uncommon (25%) in comparison with a subjective rendering in the form of an ***experience*** (53%) or an ***opinion*** (34%) (Figure 4). The other common modes of presentation are ***questions*** (9%) and ***answers*** (16%), which often occur as adjacency pairs. We also found a few ***other*** uncommon modes (7%) such as recommendations or hypotheses based on specific knowledge.

2.2 Implications for dialogue agents

The four aspects – acknowledgement, transition, detail selection and presentation – are essential ingredients and indicative of quality conversation. They provide us with finer-grained questions amenable to human evaluation: *"How does the agent acknowledge?"*, *"Was it a smooth transition?"*, *"Does the utterance contain the right level of detail?"*, and *"Was the information presented as experience or an opinion?"*.

These four aspects are also more actionable than the evaluation metrics used in prior work. They can inspire new techniques that are purposefully built to emulate these strategies. For instance, transitions can be improved with purpose-built information retrieval methods that use commonalities and differences to choose a new topic. To improve detail-selection, an agent could keep track of user knowledge and pragmatically select the right level of detail. Moreover, in their datasets, Dinan et al. (2019) and Gopalakrishnan et al. (2019) asked people to reply using knowledge snippets, but that can lead to factual statements dominating the presentation strategies. We hope that newer datasets either suggest ways to reduce this bias or not provide knowledge snippets to humans in the first place but instead post facto match utterances to knowledge snippets.

In the rest of the paper, we focus on generating responses with better acknowledgements. This is because current neural generation methods perform poorly in this regard when compared with the other aspects. They fail to acknowledge prior turns and even when they do, the acknowledgements are shallow and generic (e.g., backchannel). We hypothesize that the bottleneck is not the modeling capacity, but rather our inability to extract acknowledgements. The responses are not specific w.r.t. conversational context, a prerequisite for richer acknowledgements (e.g., shared experiences). We show that selecting responses specific to conversational context improves acknowledgements and overall quality. More broadly, we are able to demonstrate the utility of our linguistic analysis in evaluating and improving a dialogue agent.

3 A method for richer acknowledgements

Current neural generation methods typically offer short and formulaic phrases as acknowledgements: "That's interesting", "I like that", "Yeah, I agree". Such phrases are appropriate almost everywhere and convey little positive evidence for understanding or grounding. The training corpus, on the other hand, contains richer acknowledgements, which generated responses should be able to emulate.

We assume that the representational capacity of current neural models is sufficient and that out of all the sampled responses, some do indeed contain a richer form of acknowledgement. We posit that non-existent or poor sample selection strategies are to blame and that without a good sample selection strategy, improvements to the dataset, model or token-wise sampling methods are unlikely to help.

We hypothesize that responses that are more specific to conversational history provide better evidence for understanding and hence contain richer acknowledgements. As a baseline sample selection strategy, we first consider maximum pointwise mutual information (Max-PMI) (as used by Zhang et al. (2020)) between the generated response and the conversational contexts (i.e., new factual content and conversational history). However, this is insufficient because it is an imprecise measure of specificity w.r.t. conversational history. Instead, we use pointwise conditional mutual information (PCMI) to maintain specificity with individual contexts and propose a combination of PMI and PCMI scores to select overall better quality responses than Max-PMI.

Conversational rephrasing The choice of new factual content is a confounding factor for analysis. Hence, we define a simplified task, *conversational rephrasing*, where content is provided as an input. Thus, conversational rephrasing is a generation task where conversational history (**h**) and new factual content (**k**) are given as inputs and a response (**g**) is generated as the output (Figure 1). We expect the generation **g** to paraphrase the new factual content **k** in a conversational manner by utilizing the conversational history **h**.

Base generator We fix the sequence-to-sequence model and token-wise sampling method and vary the sample selection strategy. The model is trained to take **h** and **k** as input and to generate **g** as the output with the language modelling loss, i.e., we minimize the token-wise negative log likelihood. During generation, tokens are sampled autoregressively from left-to-right. While sampling each token, the probability distribution is truncated using nucleus sampling (Holtzman et al., 2020) but the truncation is kept to a minimum with a high value of p for top-p sampling. Multiple diverse candidates are sampled from the base generator and now the best candidate needs to be selected.

Response	**pmi**$(\mathbf{g};\mathbf{h},\mathbf{k})$	**pmi**$(\mathbf{g};\mathbf{h})$	**pcmi**$_h$
$\mathbf{g}_1$	87	18	14
$\mathbf{g}_2$	150	18	4

Table 1: Measures of mutual information for generated responses from Figure 1. $\mathbf{g}_2$ largely copies **k**, has high $\mathrm{pmi}(\mathbf{g};\mathbf{h},\mathbf{k})$ and would be chosen by Max-PMI. $\mathbf{g}_1$'s first sentence acknowledges using **h** and bridges to **k**; it would be chosen by Fused-PCMI on the basis of pcmi_h. $\mathrm{pmi}(\mathbf{g};\mathbf{h})$ cannot differentiate the two.

PMI for overall specificity Li et al. (2016) suggest selecting the response with maximum PMI (referred to as MMI in their work) to maintain specificity and get rid of bland or low-quality samples. Pointwise Mutual Information (PMI) between two events (x, y) is a measure of change in the probability of one event x, given another event y: $\mathrm{pmi}(x;y) \equiv \log \frac{p(x|y)}{p(x)}$. We use pmi to determine the increase in likelihood of **g**, given **h** and **k**.

$$\mathrm{pmi}(\mathbf{g};\mathbf{h},\mathbf{k}) = \log \frac{p(\mathbf{g}|\mathbf{h},\mathbf{k})}{p(\mathbf{g})}$$

A candidate generation **g** with higher PMI is more likely given the two contexts **h** and **k** than otherwise and is therefore considered more specific to the contexts. A low PMI value for a candidate response implies non-specificity to either context providing a clear signal for discarding it. A high PMI is necessary but not sufficient for a candidate to be specific to both the contexts simultaneously, since mutual information could come from either context. For example, $\mathbf{g}_2$ (Figure 1) merely copies **k** but gets a high PMI score (Table 1). Whereas $\mathbf{g}_1$ acknowledges prior turn and uses **k** but gets a lower PMI score.

PCMI for contextual specificity Pointwise Conditional Mutual Information (PCMI) considers a third variable (z) and removes information due to z from $\mathrm{pmi}(x;y,z)$ to keep only the information uniquely attributable to y.

$$\mathrm{pcmi}(x;y|z) = \mathrm{pmi}(x;y,z) - \mathrm{pmi}(x;z)$$

We propose using pcmi for contextual specificity, i.e., $\mathrm{pcmi}_h = \mathrm{pcmi}(\mathbf{g};\mathbf{h}|\mathbf{k})$ for specificity

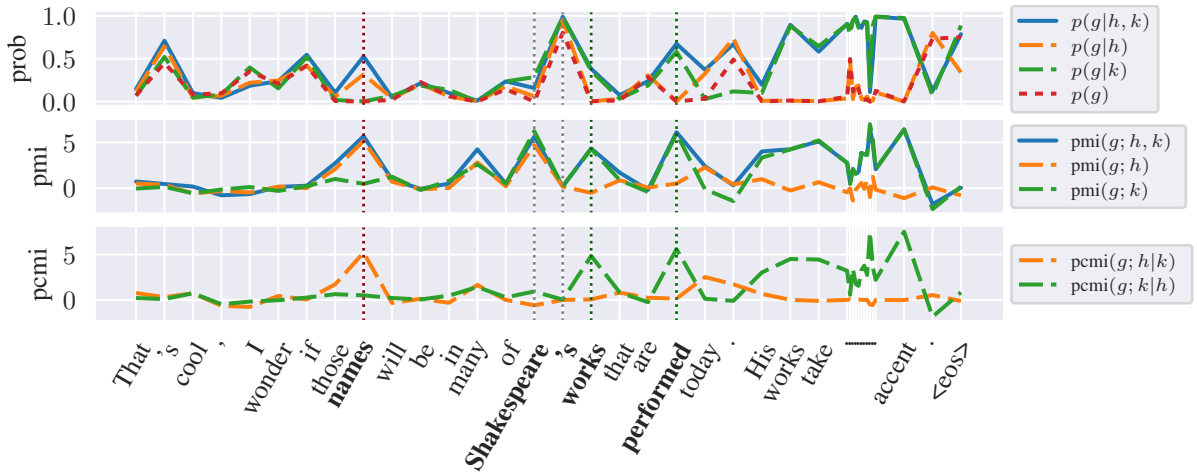

Figure 5: **Token-wise probabilities (top), pmi (middle) and pcmi (bottom) scores for the generated response g from Figure 1.** The pcmi graph is computed from the pmi graph which in turn is computed from the probability graph. The probabilities by themselves are unreliable measures of contextual specificity; the tokens predictable without **h**, **k** (e.g., *'s*) have high probability but low pmi. pmi cannot differentiate between the two contexts; tokens coming from both contexts (e.g., *Shakespeare*) have high pmi but low pcmi. pcmi differentiates the two contexts; tokens unique to conversational history **h** (e.g., *names*, *today*) have high pcmi_h, Tokens unique to new factual content **k** (e.g., *works*, *performed*, all of last sentence) have high pcmi_k.

w.r.t. to conversational history **h**, and $\text{pcmi}_k = \text{pcmi}(\mathbf{g};\mathbf{k}|\mathbf{h})$ for specificity w.r.t. new factual content **k**.

Since acknowledgement strategies are primarily based on the history of the conversation thus far, we would expect candidates with higher pcmi_h *to exhibit more human-like acknowledgement strategies.*

As a point of comparison, consider using $\text{pmi}(\mathbf{g};\mathbf{h})$ instead of pcmi_h. In our setting of conversational rephrasing for informative dialogue, **k** topically overlaps with **h**. If **g** merely copied over the new factual content **k** without any reference to **h**, it would still have a high $\text{pmi}(\mathbf{g};\mathbf{h})$ due to topical overlap but a low pcmi_h. Going back to Table 1, we can see that $\text{pmi}(\mathbf{g};\mathbf{h})$ is unable to distinguish between the two examples but pcmi_h is.

In Figure 5, the above quantities are broken down to token-level granularity. We can see that specific words that are uniquely attributable to each context are cleanly separated by both pcmi_h and pcmi_k.

Combining PMI & PCMI for overall quality To show the utility of pcmi_h in improving overall quality, we propose a heuristic method to find a more balanced response (**Fused-PCMI**) than the Max-PMI response. *For every Max-PMI response with a low* pcmi_h*, we consider an alternative that has both high* pcmi_h *and an acceptable PMI.* If such an alternative is found, we select that as the Fused-PCMI response; otherwise we default to the Max-PMI response as the Fused-PCMI response. We consider a PMI score in the top 50% of the candidate set as acceptable. To compute pcmi thresholds, we calculate quantiles based on the entire validation set and consider pcmi_h in the first quartile to be low and pcmi_h in the fourth quartile to be high. This approach is less susceptible to outliers, more interpretable and easier to calibrate than a weighted arithmetic or geometric mean.

4 Evaluation Setup

We derive the data for our conversational rephrasing task from the Topical Chat dataset (Gopalakrishnan et al., 2019). We use it to fine-tune a large pre-trained neural language model. This forms the base model as described in Section 3. To evaluate our proposed methods, we design three experiments and perform a comparative study with human annotators.

Topical Chat Dataset This is a human-human chat dataset where crowd-workers were asked to chat with each other around certain topics. They were provided with relevant interesting facts from the "Today I learned" (TIL) subreddit which they could use during the conversation. TILs are are short (1–3 sentences), self-contained, interesting facts, most of them from Wikipedia articles. When an utterance can be matched to a TIL (based on a TF-IDF threshold of 0.12), we create an instance for the conversational rephrasing task: with the utterance as **g**, the two previous utterances as **h** and

the corresponding TIL as **k**. We split the instances into training, validation and test sets (sizes in Section A.1) such that all utterances related an entity belong to the same set.

Base Model We use the GPT2-medium model (24-layer; 345M params) pretrained on the English WebText dataset (Radford et al., 2019), as implemented in HuggingFace's TransferTransfo (Wolf et al., 2019b,a) framework. Fine-tuning is performed using the language modelling objective on the training set with default hyperparameters until lowest perplexity is reached on the validation set. During generation, we sample tokens using nucleus sampling (Holtzman et al., 2020) with $p = 0.9$ and temperature $\tau = 0.9$ and get candidate responses. To compute auxiliary probabilities $\{p(\mathbf{g}|\mathbf{h}), p(\mathbf{g}|\mathbf{k}), p(\mathbf{g})\}$ for these candidates, we use separate ablation models. The ablation models are trained similar to the base model but after removing respective contexts from the training inputs.

4.1 Experimental Design

To validate our proposed methods, we do a paired comparison (on Amazon Mechanical Turk) where human annotators are shown two prior turns of conversational history and asked to choose between two candidate responses. Annotators are allowed to mark both candidates as nonsensical if the responses don't make sense. In Section A.3, we show the interfaces used to collect annotations from Amazon Mechanical Turk. Each pair of responses was compared by three annotators – we consider a candidate to be better than the other when at least two of them (majority) agree upon it. For each of the following three experiments, we compare 100 pairs of candidates generated using instances from the test set. The null hypothesis (H_0) for the three experiments is that there is no difference between the methods used to generate the candidates and we hope to reject the null hypothesis in favor of the alternate hypothesis (H_1) at a significance level (α) of 0.05.

Exp 1: PMI and overall quality First, we want to confirm that *high PMI responses are overall better quality than randomly chosen candidates* (H_1). To do so, we first generate 10 responses for each instance and compare the response having maximum $\text{pmi}(\mathbf{g}; \mathbf{h}, \mathbf{k})$ (Max-PMI) with a randomly chosen response from the remaining 9. We ask human annotators to pick the overall better candidate response.

Exp 2: pcmi_h and acknowledgement We test if *responses having high* pcmi_h *provide better acknowledgement* (H_1). To do so, we first sample 100 responses (larger than previous experiment) and out of all possible pairs keep those with $|\Delta\,\text{pcmi}_h| > 15$ (larger than population interquartile range; Figure 8). To control for the amount of new information being added, we pick pairs with closest values of pcmi_k (recall that pcmi_k denotes information uniquely attributable to **k**). Such selected pairs have $\text{Median}|\Delta\,\text{pcmi}_k| = 0.42$. We ask annotators to pick the response that provides better acknowledgement and select an acknowlegement span to support their claim.

Exp 3: Fused-PCMI vs. Max-PMI We test if *the proposed method, Fused-PCMI (that combines PMI and PCMI) selects better responses than Max-PMI* (H_1). For Fused-PCMI, we set low and high pcmi_h thresholds to be 5 and 14 respectively based on population quartiles. For instances where the Fused-PCMI response is different from the Max-PMI response, we compare the two. We consider 10 candidate responses for each test instance and find that for around 10% of the instances the Fused-PCMI candidate is different from the Max-PMI candidate. Human annotators are then asked to pick the overall better response of the two.

5 Results & Analyses

Based on human annotations, we are able to reject H_0 in favor of H_1 in all three experiments (Table 2)[3]: high PMI responses are overall better quality than randomly chosen candidates, responses having high pcmi_h provide better acknowledgement, and Fused-PCMI selects better responses than Max-PMI.

Exp	n	K	p	κ
1	87	55 (63%)	0.009	0.18
2	95	70 (74%)	3e−6	0.48
3	99	59 (60%)	0.035	0.11

Table 2: **Human annotation results.** Out of 100 instances, majority agreement was reached in **n** instances. The majority rejects the null-hypothesis (H_0) in favor of the alternate hypothesis (H_1) in **K** instances. **p** denotes the p-value and κ denotes Fliess kappa for Inter-annotator agreement.

While according to Exp 1, high PMI responses are overall better quality, upon further analysis

[3] Statistically significant with $p < 0.05$ (Binomial Test).

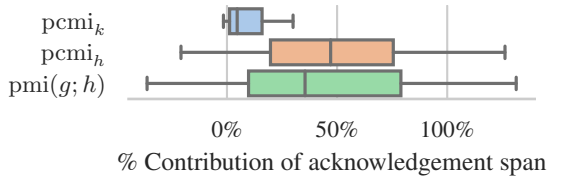

Figure 6: **Attribution to acknowledgement span.** A larger fraction of pcmi_h can be attributed to human annotated acknowledgement spans compared to $\text{pmi}(g;h)$ and pcmi_k.

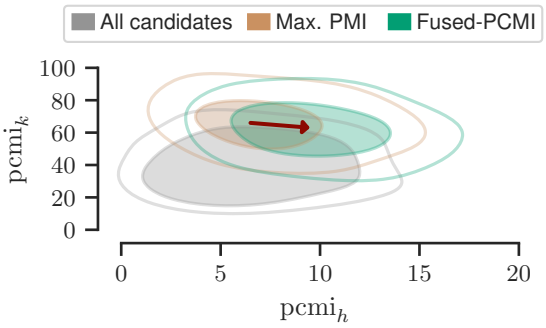

Figure 7: **Bivariate Kernel Density Estimate plot All candidates, Max-PMI responses and Fused-PCMI responses.** Bivariate kernel density estimate plot w.r.t. $\textbf{pcmi}_k$ and $\textbf{pcmi}_h$ at levels 0.5 and 0.75. We see that Fused-PCMI responses compared with Max-PMI trade off little pcmi_k for a large relative gain in pcmi_h. See Figure 8 in Section A.2 for univariate box plots.

we find that *PMI is useful for filtering out bad samples, but not necessarily for selecting between the good samples.* When paired with a random response from the top 50% of the candidates (ranked according to their PMI), people prefer the Max-PMI response only 52% of the time (not significant). On the other hand, if the random response was in the bottom 50%, then the Max-PMI response is preferred 74% of the time.[3]

In Exp 2, we ask annotators to mark text-spans that indicate acknowledgement (Table 3). If token-level pcmi_h is concentrated in these spans, we have further proof that pcmi_h indicates acknowledgement. Indeed, in Figure 6, we see that pcmi_h is most attributable to the acknowledgement spans, followed by $\text{pmi}(g;h)$ and pcmi_k. *Thus,* pcmi_h *captures acknowledgements with greater specificty than* $\text{pmi}(g;h)$.

To understand the mechanism behind the improvement in Exp 3, we look at the distribution of samples w.r.t. pcmi_k and pcmi_h in Figure 7. We observe that Max-PMI responses heavily skew the distribution towards higher pcmi_k, whereas Fused-PCMI responses show a more balanced improvement along both pcmi_h and pcmi_k. *Fused-PCMI increases both* pcmi_h *and* pcmi_k *(medians cross 75% quartiles), indicating that the responses are simultaneously specific to both* ***h*** *and* ***k****.*

6 Discussion

We show that samples with higher pcmi_h provide better acknowledgement and Fused-PCMI improves overall quality compared to Max-PMI. Thus, by improving acknowledgements – an aspect we identified during our analysis of human strategies – we were able to improve overall quality. This demonstrates the utility of linguistic analysis for finding interpretable and actionable metrics.

While we show that our learnings apply to informative dialogue which adds factual knowledge (Dinan et al., 2019; Parthasarathi and Pineau, 2018; Gopalakrishnan et al., 2019), we expect it to generalize to any dialog setting that adds new content, e.g., experiences (Ghazvininejad et al., 2018) and personas (Zhang et al., 2018). Any dual-context language generation task where the two contexts are asymmetric in their information content can potentially benefit from PCMI.

There is scope for improvement: Max-PMI still selects better responses than Fused-PCMI in 40% of the instances. This could be because it is easy for the model to copy over **k** and generate a high PMI response that is also fluent and accurate. Fused-PCMI encourages synthesis of acknowledgement using **h** and abstraction over **k** and it could therefore be prone to disfluencies and inaccuracies. We hope that orthogonal modeling improvements (Meng et al., 2020) reduce such effects.

A cause for concern with the human evaluation is low inter-annotator agreement for Exp 1 and 3 where we ask them to pick responses with "overall better quality and suitability". However, quality measurements are inherently subjective; people differ in the importance they place on different aspects such as engagement, informativeness, fluency etc., as corroborated by prior work (Finch and Choi, 2020) that shows low Cohen's kappa (0.13, 0.22) for overall quality judgements. In this work, diverse expectations from multiple annotators are captured yet subsequently averaged into "overall quality". We leave it to future work to find finer-grained metrics that have high inter-annotator agreement and derive empirical weights to combine them into "overall quality".

In this work, we looked at acknowledgements and their impact on quality in isolation, but in a real

Previous Utterance	High pcmi_h candidate
No I did not! Han Solo was apparently also an imperial lieutenant before meeting up with Chewbacca.	**that is very interesting and I wonder if he was one of the first or the first one to meet Chewba.** I was just reading that George Lucas originally intended Han to be a green alien
I did not. That sounds interesting. Much like the "tip of the hat" originates from Knights raising their visors.	**Yes, that sounds like a very cool tradition.** I wonder how they enforced that law, it sounds as ridiculous as the prohibition of wearing a hat in 1912, that was a law that would prevent the use of hat pins by suffragettes.
Yeah. When Apollo 11 was preparing for their voyage, they had discussion as to whether they should plant a United Nations flag instead of a US flag. If I planted a UN flar and someone asked me "whose flag is that?" I would answer, mine…	lol, no kidding. **I wonder if they are still discussing the issue with the moon?** I can't believe the moon is moving away from us at about the rate that your nails grow. That seems really slow to me.

Table 3: Human annotated spans of text that indicate acknowledgement (in bold) in candidate responses with higher $\textbf{pcmi}_h$ from Exp 2.

	Previous Utterance	Max-PMI candidate	Fused-PCMI candidate
1	I live in the southwest USA desert. Not much in the way of water here either, but we've got some nice artificial pools. What kind of swimming did you do?	There is a type of competitive swimming where the athlete has their hands and legs bound, and swims like a dolphin. I have never heard that before. It seems fishy to me. I'm not sure I can swim like that.	I went competitively but was having a tough time. I liked that kind of competition where the swimmer can swim with their hands bound.
2	It is interesting. It's also interesting that they are trying to plan a Mars colonization project that will be funded by a reality show, I really want to see that!	A satellite was lost on Mars because of the use of the Imperial system and NASA used the metric system.	Sounds like a weird reality show. I'm going to look it up. A satellite was lost on Mars because it was used by NASA using the metric system.
3	Yeah I definitely like clothes. I think it's interesting how many uses hats have, like for protection, ceremonial reasons, religious reasons, safety, and fashion.	Did you know that until 1998, British politicians had to wear a top hat if making a point of order in the house of commons?	In 1998, British politicians were required to wear a top hat if making a point of order in the house of commons. I guess that's kind of messed up and counterproductive.
4	Yeah it's super funny that he had a rap battle in the lunchroom. Seems like something straight out of a movie!	Ha, I agree. Speaking of rappers, did you know that "California Gurls" was actually a West Coast vs East Coast answer song to Alicia Keys and Jay Z's "Empire State of Mind"	Yeah, and how about the fact that the California Gurls song was actually a west coast vs east coast answer song to a pair of famous rappers. That is surprising
5	im not surprised. rats are quite clever, and i guess if scientists understand the rat brain, then they can wire up the interface. cool.	It's amazing how they can be clever and change forms. I like how they can now Debrate a chicken with surgical precision.	Robots can also debone chicken with precision. I guess that would help avoid chicken pogis.

Table 4: **Illustrative samples of selected responses used in Exp 3.** For samples 1,2 and 3 people prefer Fused-PCMI and for samples 4 and 5 they prefer Max-PMI. Factual content copied largely verbatim by the model is in gray. Specifically, the Fused-PCMI candidate in 1 answers the question (Max-PMI does not) and in 3 refers back to contradict utility of hats.

system, the performance of the model also depends on other factors like user compliance and the retrieval model. In practice, we think the interplay between the four linguistic aspects is critical and needs to be explored. For instance, preliminary experiments with live conversations and an off-the-shelf retriever suggested that a bad choice of **k** with tenuous connections to **h** can make synthesis harder and lead to lower quality Fused-PCMI responses. Better retrieval models (Ren et al., 2020) that make use of transition strategies to determine **k** can lead to better acknowledgements.

In this work, we identified salient aspects of human-human informative conversations and found deficiencies in current neural dialogue systems. We proposed a PCMI-based selection strategy that selected responses with acknowledgements and higher overall quality. We hope that our work provides actionable insights and metrics for future work and more generally inspires the use of linguistic literature for grounding conversational research.

7 Acknowledgements

We are grateful to Amelia Hardy, Nandita Bhaskhar, Omar Khattab, Kaitlyn Zhou, Abigail See, other Stanford NLP group members and the anonymous

reviewers for helpful comments. This research is funded in part by Samsung Electronics Co., Ltd. and in part by DARPA CwC under ARO prime contract no. W911NF-15-1-0462. This article solely reflects the opinions and conclusions of its authors. Christopher Manning is a CIFAR Fellow.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.

Herbert H Clark and S Haviland. 1977. Comprehension and the given-new contract. In Roy O Freedle, editor, *Discourse production and comprehension*, pages 1–40. Lawrence Erelbaum Associates, Hillsdale, NJ.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, Shui Hu, Karthik Gopalakrishnan, Seokhwan Kim, Lauren Stubel, Arindam Mandal, and Dilek Hakkani-Tür. 2020. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize Proceedings*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Herbert P Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech acts*, pages 41–58. Brill.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ellen A Isaacs and Herbert H Clark. 1987. References in conversation between experts and novices. *Journal of experimental psychology: general*, 116(1):26.

Robert M Krauss and Susan R Fussell. 1996. Social psychological models of interpersonal communication. *Social psychology: Handbook of basic principles*, pages 655–701.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Refnet: A reference-aware network for background based conversation. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, USA.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume abs/1908.09528.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Lui, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, Y-Lan Boureau, and Jason Weson. 2020. Recipes for building an open-domain chatbot. ArXiv preprint arXiv:2004.13637.

Harvey Sacks and Gail Jefferson. 1995. *Winter 1971*, chapter 12. John Wiley & Sons, Ltd.

Vicki L Smith and Herbert H Clark. 1993. On the course of answering questions. *Journal of memory and language*, 32(1):25–38.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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 *ACL, system demonstration*.

A Appendix

A.1 Model training details

Each model (main and ablation) was trained on a single NVIDIA Titan Xp GPU for 5 epochs and took approximately 8 hours to train. The training dataset had 51407 instances, validation 2491 and test 2728. The Topical Chat dataset and Switchboard corpus are in English language. The main model used for response generation had a validation loss (average negative log liklihood) of 2.05 which it reached after 2 epochs.

A.2 Univariate distribution

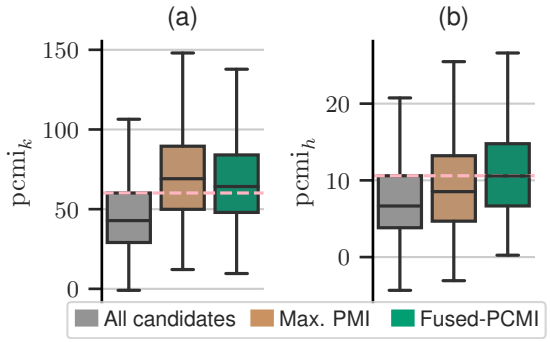

Figure 8: **Univariate box plots for All candidates, Max-PMI responses and Fused-PCMI responses.** (a) is w.r.t. pcmi_k and (b) w.r.t pcmi_h. Pink horizontal lines indicate 75% quartile for All candidates. Max-PMI responses (orange) have high pcmi_k (median above pink line), but low pcmi_h. Fused-PCMI responses (green) show balanced yet high pcmi_h and pcmi_k (medians cross pink lines).

A.3 Annotation Details

We had 9, 19 and 19 unique annotators for experiments 1, 2 and 3 respectively. All three annotators agreed in 32/87 instances for experiment 1, 52/87 instances for experiment 2 and 32/99 instances for experiment 3.

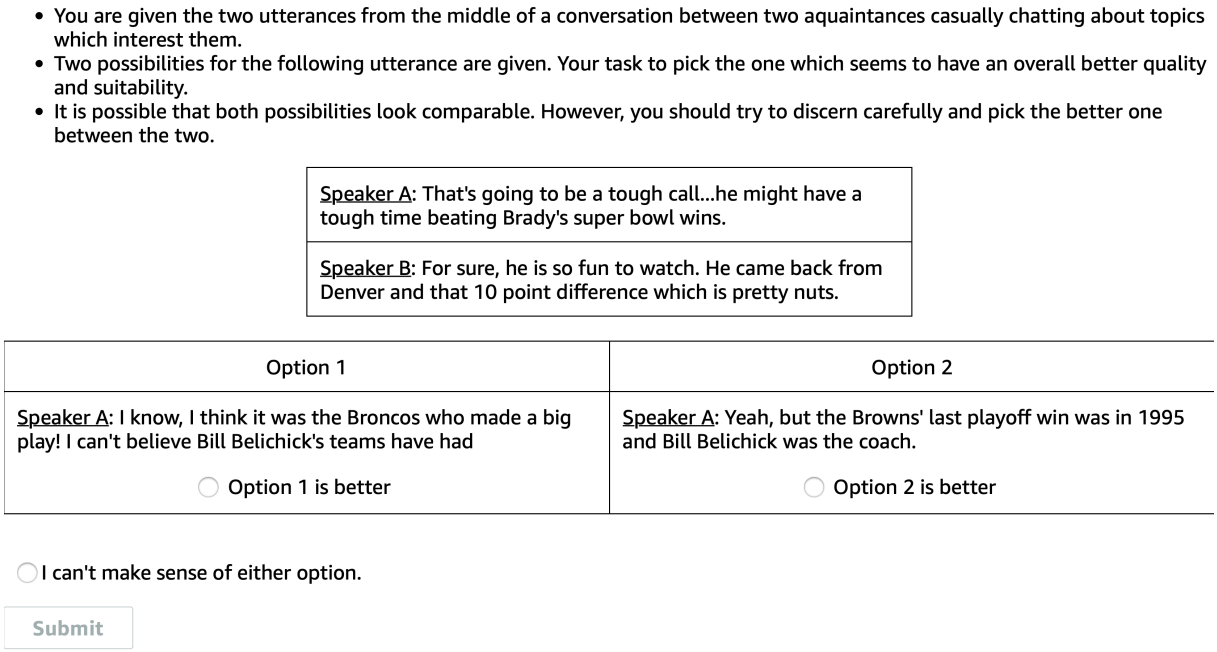

Figure 9: Annotation interface for Best PMI v/s rest

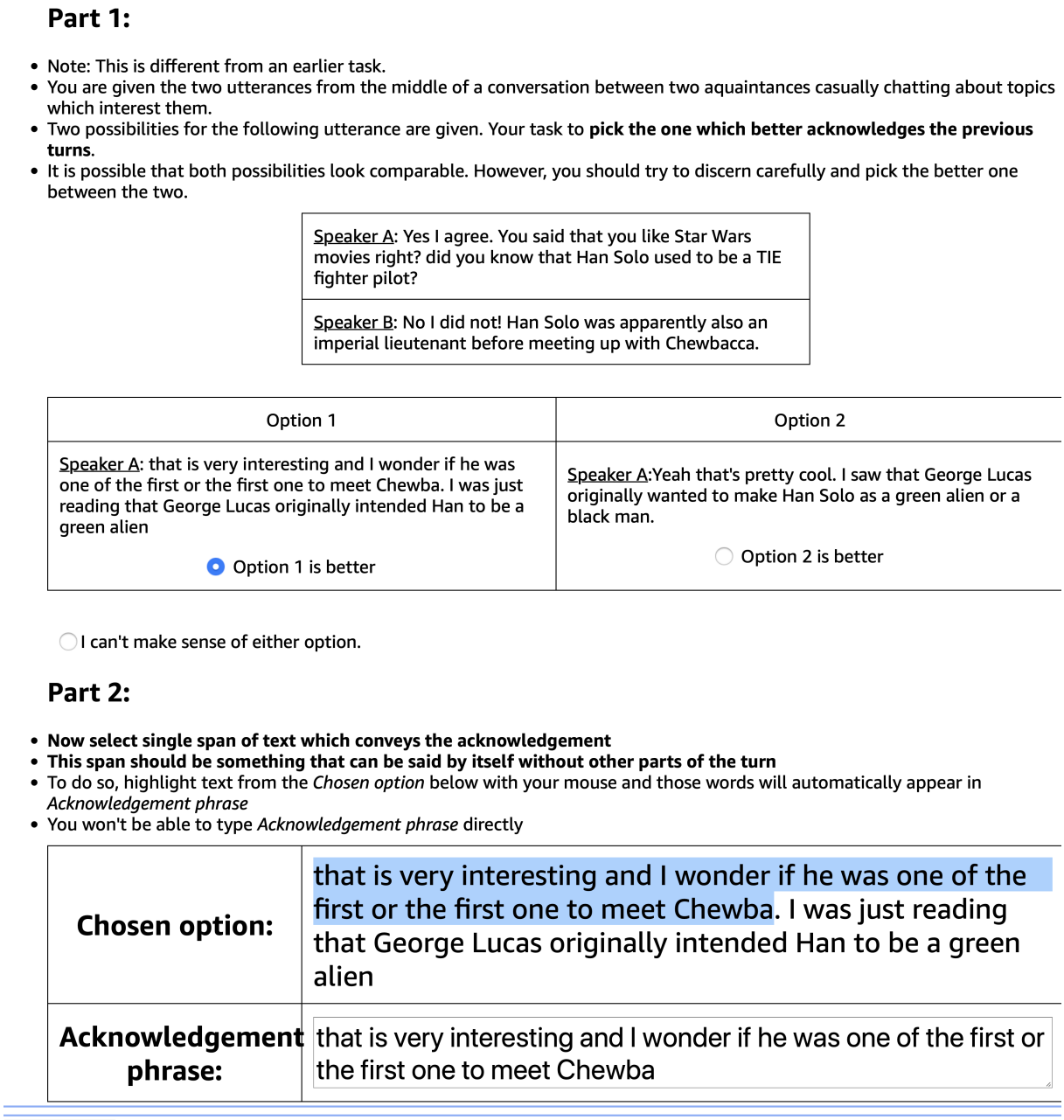

Figure 10: Annotation interface for acknowledgement differences due to pcmi_h