Human-like informative conversations via conditional mutual information

Ashwin Paranjape
Stanford University
ashwinp@cs.stanford.edu

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

Abstract

The goal of this work is 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, examining how humans apply strategies for *acknowledgement*, *transition*, *detail selection* and *presentation*. However, when current chatbots (explicitly provided with new factual content) introduce facts in a conversation, their generated responses do not ***acknowledge*** the prior turns. This is because, while current methods are trained with two contexts, new factual content and conversational history, we show that their generated responses are not simultaneously specific to both the contexts and in particular, lack specificity w.r.t. conversational history. We propose using *pointwise conditional mutual information* (pcmi_h) to measure specificity w.r.t. conversational history. We show that responses that have a higher pcmi_h are judged by human evaluators to be better at acknowledgement 74% of the time. To show its utility in improving overall quality, we compare baseline responses that maximize *pointwise mutual information* (Max. PMI) with our alternative responses (Fused-PCMI) that trade off pmi for pcmi_h and find that human evaluators prefer Fused-PCMI 60% of the time.

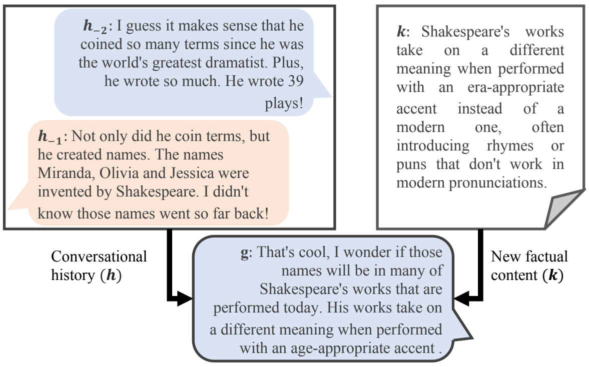

h_{-2}: I guess it makes sense that he coined so many terms since he was the world's greatest dramatist. Plus, he wrote so much. He wrote 39 plays!
h_{-1}: Not only did he coin terms, but he created names. The names Miranda, Olivia and Jessica were invented by Shakespeare. I didn't know those names went so far back!
k: Shakespeare's works take on a different meaning when performed with an era-appropriate accent instead of a modern one, often introducing rhymes or puns that don't work in modern pronunciations.
Conversational history (h)
New factual content (k)
g: That's cool, I wonder if those names will be in many of Shakespeare's works that are performed today. His works take on a different meaning when performed with an age-appropriate accent .

Figure 1: Conversational history ($\mathbf{h} = (h_{-2}, h_{-1})$) and new factual content (**k**), two largely independent contexts, are used to generate the response (**g**).

1 Introduction

Social chatbots are becoming viable and are being widely deployed to converse with humans (Gabriel et al., 2020). In part, this progress is driven by advances in neural generation (Adiwardana et al., 2020; Roller et al., 2020) which can handle a wide variety of user responses and 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, a principle for human conversations codified in the *given-new* contract (Clark and Haviland, 1977).

Past work has used neural-generated bot responses for adding new world knowledge (Dinan et al., 2019; Gopalakrishnan et al., 2019), reviews (Ghazvininejad et al., 2018) and personality (Zhang et al., 2018) into conversations. However, the lack of control over neural generation methods makes it difficult to reliably use them to introduce new information. Furthermore, prior work uses coarse evaluation metrics such as “engagingness”, “appropriateness” and “informativeness”, that don’t capture fine-grained traits of human conversations and hide deficiencies in generated responses. Psycholinguists and sociolinguists, on the other hand, have studied human conversations in a variety of settings and have identified defining conventions, principles and contracts (Grice, 1975; Clark, 1996; Krauss and Fussell, 1996) which can guide ML models and evaluation approaches.

Our first contribution is a detailed analysis of how human conversations incorporate world knowledge from a linguistic perspective. We manually annotate conversations from the Switchboard corpus to identify key traits indicative of how humans incorporate world knowledge in conversations. 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.

We would like these strategies to be emulated by machine-learned models attempting to incorporate new factual content into conversations. We consider a simplified task of **conversational rephrasing** (see Figure 1), in which the factual content to be added is not left latent but is provided as a text input to the model, 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).

Our second contribution is using information theoretic metrics to select generated responses that exhibit human-like acknowledgement. Fine-tuning a large pre-trained language model and sampling responses from it with low truncation decoding (higher top-k, top-p (Holtzman et al., 2020)) can provide good generations but their quality is highly variable. We observe that while maximum pointwise mutual information (Max. PMI) used in prior work (Li et al., 2016; Zhang et al., 2020) filters out responses that are not specific enough, a generated response that simply copies over the new factual content while largely ignoring the conversational history can still have high mutual information (MI) with the input.

To quantify the amount of information drawn from each context (i.e., 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. For 10% of the evaluated instances, we are able to find such alternative candidates and find that despite a lower PMI, human annotators prefer Fused-PCMI alternative over the Max. PMI response 60% of the time.[1] We release[2] annotated conversations from the Switchboard corpus, code for fine-tuning and calculating scores along with human evaluations of proposed methods.

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 approaches to conversational analysis. According to his *given-new* contract (Clark and Haviland, 1977), the speaker tries to connect their utterances with the given information (assumed to be known to the listener) and add 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 by step, connecting the given, stated information to new information. We annotate these as different transition strategies.

Detail selection strategies According to Isaacs and Clark (1987), the speakers 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 and classify utterances as such.

Presentation strategies According to Smith and Clark (1993), the presentation of answers 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. opinion, experience, recommendation) which we classify into various presentation strategies.

Our eventual goal in this paper is to build a system that conversationally paraphrases facts. However, in real life conversations, new information can come from a variety of sources, not all of them relevant to the task. Thus, for the purposes of this annotation, we divide new knowledge into *general*, *specific*

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

[2] url

Strategy	Example	
Agreement	Prev:	Well, I think they are a lot better at making movies than they used to
	Reply:	The quality I think maybe has improved in that respect . . .
Shared Experience	Prev:	I am more interested in watching some of the movies that are on TV.
	Reply:	Well, that's probably what I watch most frequently the movies . . .
Backchannel	Prev:	There is a lot of places in the United States I still want to go to.
	Reply:	*Uh huh, yeah.* Now, have you been to Yellowstone? . . .

Table 1: Some **Acknowledgement strategies** from Switchboard. Parts of the turns have been omitted for brevity.

and *experiential*, similar to the division of common ground into personal and communal (Clark, 2006). **General knowledge** is expected to be known to the population as a whole, whereas **experiential knowledge** can only be derived from embodied experiences. We categorize all other knowledge as **specific knowledge**, which is knowledge that can be "looked up" but isn't widely known. We are interested only in specific knowledge in this work as it is a source of new information in a conversation and is also likely to be available as text for machine-learned systems.

2.1 Analysis of strategies

Dataset We annotate part of the The Switchboard Dialog Act Corpus (Stolcke et al., 2000) which extends 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, the conversations are more intimate and realistic than text-based conversations with 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.

First, we look at the prevalence of **specific knowledge** (as defined in the previous section) in our annotated corpus. Out of 408 annotated turns, 111 (27%) incorporate specific knowledge. If we look at whitespace-separated tokens, turns incorporating specific knowledge account for 56% of overall tokens. Next, we analyze various strategies employed in turns containing specific knowledge:

Acknowledgement Strategies We find three major categories of acknowledgement strategies: ***agreement*** (or disagreement), ***shared experiences*** (or differing experience) and ***backchanneling*** (Table 1). Together, these account for around 60% of the turns (Figure 2), while around 30% of the turns offer no acknowledgement corroborating Clark and Brennan (1991). In certain cases such as answering a question, the answer itself is an acknowledgement of understanding the question and thus any explicit acknowledgement isn't necessary (categorized as ***N/A***).

Transition Strategies New information added to a conversation is semantically connected to the previous turns using a transition strategy (Table 2; Figure 2). The ***discussion theme*** is typically used (13%) at the beginning of conversations, or as an anchor point after a certain conversational direction has been exhausted. Using ***commonalities*** and ***differences*** is a more prevalent strategy (28%) in the middle of a discussion when a person wants to steer the conversation in a particular direction. In other cases, a speaker may elaborate their own turn (***self-elaboration***, 32%) with a supportive listener or may elaborate the other person's turn (***other-elaboration***, 22%) to signal interest.

Detail-selection strategies We find that people probe the other speaker's knowledge about an entity before diving into details about it. Around 50% of times, an entity is introduced just by name and without any details (***introduce entity***) compared to 66% of times when the details are laid out (***details***). Note that a turn can have both labels, i.e. it can introduce an entity for the first time, can have details about it and also introduce another entity. Interestingly, sometimes an entity's name is omitted (7%), either as an abstraction or because they can't recall, creating an opening for the other speaker to chime in.

Presentation strategies A single utterance can have multiple modes of presentation. In Figure 2, we can see their frequencies in our annotated corpus. Surprisingly, a ***factual*** (objective) statement of specific knowledge occurs only 25% of times,

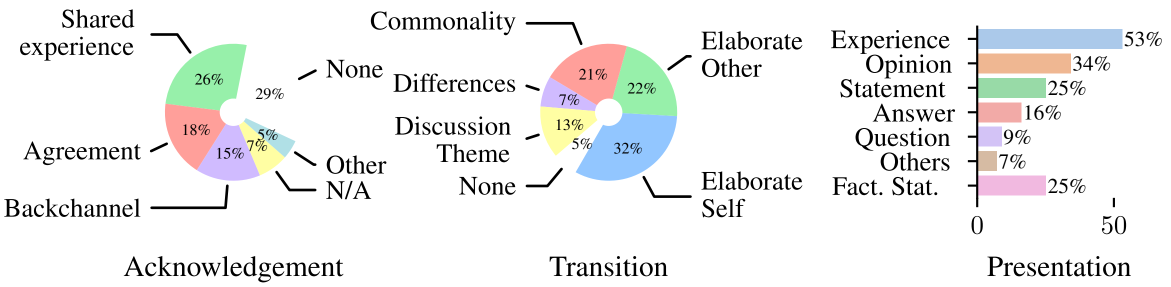


Figure 2: Distribution of acknowledgement, transition and presentation strategies

Strategy	**Example**	
Commonality (with new topic)	Prev: Reply:	... interested in watching some of the movies that are on T V ... ... like nostalgic older movies ... like *the MARX BROTHERS* ... .
Differences (with new topic)	Prev: Reply:	... I go from classical all the way to, uh, jazz and country ... ... certain types of country western I can't handle *that twangy stuff*,
Elaborate other (same topic)	Prev: Reply:	And it was shocking at the end too. So, - Absolutely. Uh, but much more true to life and I think that is, the point.
Elaborate self (same topic)	Self: Prev: Reply:	how to build things and, um, they have a calligraphy show, I watch that. Oh, that's nice. And, um, they have a lot of cooking shows, And, oh, you know ...
Discussion Theme	Prev: Reply:	... But other than that, I like pretty much everything. so, other than, uh, - as far as instruments, I can go from piano to the ...

Table 2: Some **Transition strategies** from Switchboard corpus. Parts of the turns have been omitted for brevity.

whereas a subjective rendering as an ***experience*** (53%) or ***opinion*** (34%) is more common. ***Questions*** (9%) and ***answers*** (16%) often occur as adjacency pairs. ***Other*** modes (7%) include specific knowledge-based recommendations or hypotheses.

Implications for dialogue agents Among the four strategies outlined above, transition involves choosing new factual content to be introduced. This is a hard task that needs methods from the field of information retrieval and we consider it beyond the scope of this work. On the other hand, based on advances in neural language generation and the authors' own experience, we find that current methods demonstrate detail selection and presentation strategies to an acceptable degree. However, neural generation methods fail to acknowledge prior turns as well as humans do, because the generated responses are not specific w.r.t conversational context. In the following section, based on this key insight, we propose methods to select generations with better acknowledgement, thus significantly improving their overall quality.

3 Methods for human-like rephrasing

In this section, we first define the task of conversational rephrasing followed by the description of our approach. We use a base model to generate candidate responses and select a single best response. The candidates are of variable quality and the goal is to pick the most human-like candidate as the best response. As a scoring function, we first consider pointwise mutual information (PMI) (as used by Zhang et al. (2020)) between the generated response and the conversational contexts (i.e. new factual content and conversational history) but show that it might not be specific to conversational history. Instead, we propose using pointwise conditional mutual information (PCMI) to maintain specificity with conversational context. As we shall see later, it also serves as a measure of acknowledgement. Finally, we propose a combination of PMI and PCMI scores to select overall better quality responses than we could have with PMI alone.

Conversational rephrasing We define conversational rephrasing as a generation task where conversational history (**h**) and new factual content (**k**) are inputs and a response (**g**) is generated as the output (as shown in Figure 1). We expect the generation **g** to paraphrase the new factual content **k** in a conversational manner by utilizing the conversational history **h**. Not including the prior task of finding the right **k** simplifies the task and analysis.

Base model As is standard, we train a sequence to sequence model to take **h** and **k** as input and generate **g** as the output with the language modelling loss, i.e. we minimize the token-wise negative log likelihood. During generation, we sample tokens autoregressively from left-to-right. While sampling each token, we truncate the probability distribution using nucleus sampling (Holtzman et al., 2020) but keep top-p very high, implying less truncation. Thus, by using the base model, we are able to generate multiple diverse candidates and our task is to now pick the best candidate.

PMI for overall specificity Language models can produce unspecific responses that may be bland and low-quality. Li et al. (2016) suggest improving their quality by selecting the response with maximum PMI (referred to as MMI in their work) to maintain specificity. 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 high PMI indicates that a candidate generation **g** is more likely given the two contexts **h** and **k** than otherwise and is therefore considered specific to the contexts. While we can discard a low PMI candidate because it is specific to neither context, we cannot necessarily conclude that high PMI is specific to both the contexts simultaneously, since mutual information could come from either context.

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

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

As a point of comparison, consider using $\mathrm{pmi}(\mathbf{g};\mathbf{h})$ instead of pcmi_h and the case where **k** and **h** have topical overlap (common in this setting). If **g** merely copied over the new factual content **k** without any reference to **h**, it would still have a high $\mathrm{pmi}(\mathbf{g};\mathbf{h})$ but a low pcmi_h. For instance, in Figure 3, the word *Shakespeare* is common to both contexts, so it has high $\mathrm{pmi}(\mathbf{g};\mathbf{h})$ but has a low pcmi_h because it is not unique to **h**.

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. Note that there may exist yet better ways of combining the two scores to pick the best candidate but we leave that for future work.

4 Evaluation Setup

We derive data for the conversational rephrasing task (as outlined above) from the Topical Chat dataset (Gopalakrishnan et al., 2019). We then 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

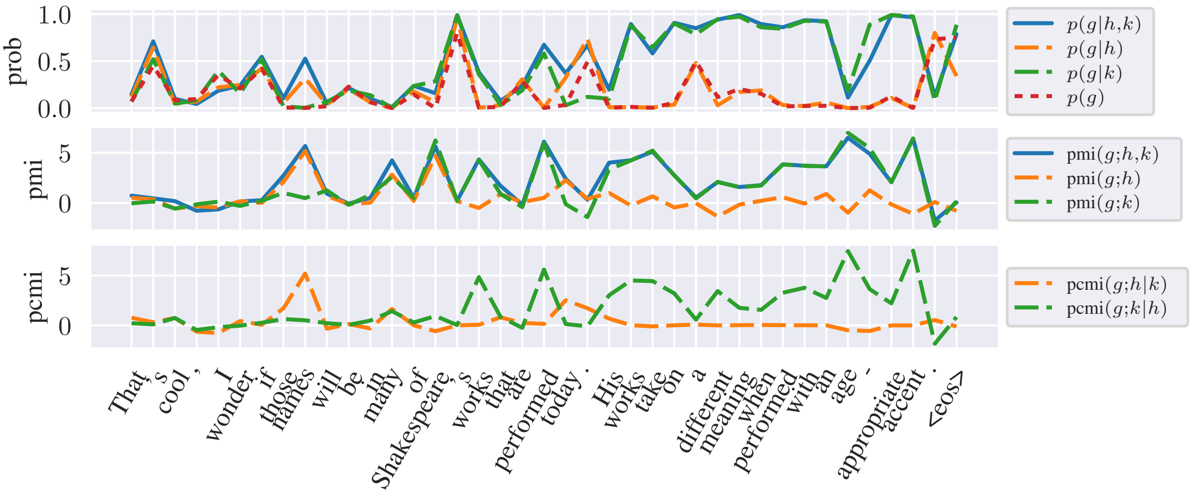


Figure 3: Token-wise probabilities (top), pmi (middle) and pcmi (bottom) scores for the generated response **g** from Figure 1. The pcmi graph is computed from the pmi graph which in turn is computed from the probability graph. Tokens predictable without **h**, **k** (e.g. *'s*), have high probability but low pmi. Tokens unique to either context (e.g. *Shakespeare*) have high pmi but low pcmi. Tokens unique to conversational history **h** (e.g. *names*, *today*) have high pcmi_h. Tokens unique to new factual content **k** (e.g all of last sentence) have high pcmi_k.

facts, most of them from Wikipedia articles. When an utterance can be matched to a TIL (based on a low 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 about 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 similarly to the base model but after removing respective contexts from the training inputs.

4.1 Experiment Design

To validate our proposed methods, we do a comparative study (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.2, we show the interfaces used to collect annotations from Amazon Mechanical Turk. Each pair of responses was compared by three annotators and we consider a candidate to be better than the other when at least two of them agree upon it. For each of the following three experiments, we compare 100 pairs of candidates generated using instances from the test set.

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

Exp 2: $\mathbf{pcmi_h}$ and acknowledgement We test if *responses having high* pcmi_h *provide better acknowledgement*. To do so, we first sample 100 responses (larger than previous experiment) and out of all possible pairs keep those with $|\Delta\mathrm{pcmi}_h| > 15$ (larger than population interquartile range; see Figure 4). 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 $\mathrm{Median}|\Delta\mathrm{pcmi}_k| = 0.42$. We ask annotators to pick the response that provides better acknowledgement and select the span of text that indicates it.

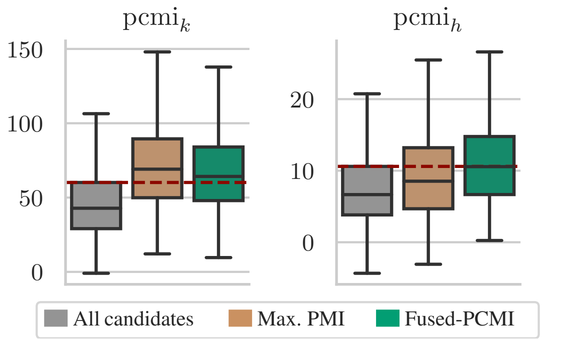


Figure 4: Distribution of pcmi_k (left) and pcmi_h (right) for All candidates, Max. PMI responses and Fused-PCMI responses. Note that they are on different scales. Red horizontal lines indicate 75% quartile for All candidates. Max. PMI responses have high pcmi_k (median above red line), but low pcmi_h. Fused-PCMI responses show balanced yet high pcmi_h and pcmi_k (medians cross red lines).

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.* 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 around 10% of the instances have Fused-PCMI response different from Max. PMI response. Human annotators are then asked to pick the overall better response of the two.

5 Results & Analysis

PMI indicates overall quality Based on human annotations, majority decision was reached on 87 out of 100 instances from Exp 1. Out of the 87 instances, the Max. PMI generation was picked by consensus in 55 instances (63.21%) over the randomly chosen response.[3] Furthermore, if PMI of the random response was in the top 50% of the candidates (ranked acccording to their PMI), then the Max. PMI response is preferred only 52% of the time (not significant). On the other hand, if it was in the bottom 50%, then the Max. PMI response is preferred 74% of the time.[3] *Thus, PMI is useful for filtering out bad samples, but not necessarily for selecting between the good samples.*

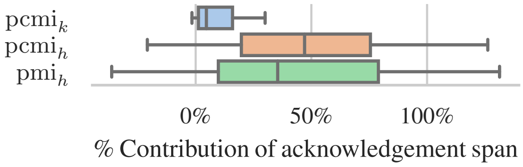


Figure 5: Human annotated text-spans indicative of acknowledgement contribute a large fraction to pcmi_k but little to pcmi_h

Max. PMI favors factual content We investigate if the Max. PMI response is specific to both contexts (**h** and **k**) simultaneously. Recall that pcmi_h and pcmi_k measure content unique to respective contexts. In Figure 4, we see that Max. PMI responses have high pcmi_k (median crosses 75% quartile) compared to all candidate responses. On the other hand, pcmi_h does not increase as much (median does not cross 75% quartile). This is because there is an assymetry between the two contexts and simply copying over content from **k** can lead to a large increase in pmi. *Thus, we conclude that Max. PMI responses are highly specific to **k** but not specific enough to **h**.*

pcmi_h indicates acknowledgement Out of 100 instances in Exp 2, majority decision was reached by human annotators for 95 instances. The candidate response with higher pcmi_h was chosen as providing better acknowledgement than the one with lower pcmi_h in 70 instances (74%).[3] Moreover, text-spans annotated by human annotators to indicate acknowledgement contribute to 47% (median) of pcmi_h, whereas they contribute only around 5% (median) of pcmi_k, as can be seen in Figure 5. *Thus, we conclude that* pcmi_h *is representative of acknowledgement and localizes it to individual tokens as well.*

Fused-PCMI is better than Max. PMI Out of 100 instances in Exp 3, majority decision was reached by human annotators for 99 instances, out of which Fused-PCMI was preferred 59 times (60%) [3]. Moreover, as can be seen in Figure 4, Fused-PCMI increases both pcmi_h and pcmi_k (medians cross 75% quartiles), indicating that the responses are specific to both **h** and **k**. *From this, we conclude that Fused-PCMI responses are specific to both contexts and are overall better compared to Max. PMI.*

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

6 Discussion & Conclusions

Prior work has proposed datasets and models for incorporating Wikipedia article snippets (Dinan et al., 2019; Parthasarathi and Pineau, 2018), foursquare tips (Ghazvininejad et al., 2018) and Reddit TILs ('Today I learnt') (Gopalakrishnan et al., 2019) into conversations. Ren et al. (2020); Meng et al. (2020) propose models to select new factual content from a given document for "background-based conversations". In this work, take a linguistic approach and identify the main strategies used by humans while including new factual content in conversations. Our focus is to make the generated responses more human-like with regards to these strategies. We find that current models provide poor acknowledgement due to lack of specificity towards conversational history. We find evidence that while PMI is good at filtering out bad responses, it biases the chosen response towards having new factual content at the expense of acknowledgement. Separately, we establish that samples with higher pcmi_h provide better acknowledgement. Finally, we propose Fused-PCMI, a heuristic method to combine pcmi_h with pmi and show that it improves overall quality compared to Max. PMI. A current limitation is that Fused-PCMI finds better alternatives over Max. PMI in only 10% of the instances because of the strict pcmi thresholds in the heuristic. There are opportunities for combining the scores in better ways, which we leave for future work.

Forward ablation models The utility of PMI in maintaining specificity of responses in dialogue was first proposed by Li et al. (2016) and recently revalidated by Zhang et al. (2020). They use a backward scoring model that computes probability of the contexts given the generation. On the other hand, in our forward ablation models, we remove contexts and compute probability of the generation. While these may lead to the same mathematical formulation, there are two advantages to using forward ablation models. Firstly, the pmi (and pcmi in our case) can be computed individually for each generated token, enabling fine-grained, span-level analysis. Secondly, the backward model conditions on human responses during training but on model-generated responses during inference (which are from two different distributions). Whereas, there is no such train-test distribution mismatch in our forward ablation models, leading to more accurate pmi and pcmi scores.

Future work based on linguistic insights In this work, while we make progress on improving the ability of current models to provide acknowledgement, insights from linguistic analyses of other strategies (Section 2) are yet to be applied. For transition strategies, methods that retrieve new factual content based on commonalities and differences would be pertinent for topical transitions, whereas methods that measure user interest and initiative would be important for choosing between self-elaboration or other-elaboration. Based on analysis of presentation strategies, dialogue agents cannot seem human-like until they express experiences and opinions as often as humans do. However, if they do so, it raises questions about the ethical responsibility of ascribing opinions and embodied experiences to virtual agents. Lastly, while our analysis of detail selection strategies provides a starting point, we believe there is a need for a deeper understanding based on linguistic principles of pragmatics. These principles can guide models to avoid both abrupt introduction of new factual content and unnecessary repetition of content that has already been introduced.

Factual correctness There is increasing concern about neural language models producing factually inaccurate generations (Kryscinski et al., 2020). While this work doesn't tackle that problem, it provides clues towards fixing it. Astute readers might notice how "era-appropriate" (from **k** in Figure 1) is replaced with "age-appropriate" (from **g** in Figure 3). We can observe a sharp dip in $p(\mathbf{g}|\mathbf{k})$ for the token "age" indicating that it is relatively unlikely compared to surrounding tokens. If a model were to compare it with "era" (a high probability counterfactual sample), it could potentially figure out that they are not equivalent.

In this work, we see that linguistic analysis can lead to improvements in neural generation methods that provide better acknowledgement. Our proposed strategy to select the best candidate response, Fused-PCMI, leads to better quality responses than Max. PMI. As discussed in this section, our linguistic analysis also lays the groundwork to improve upon other strategies for more human-like informative conversations.

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. 2006. Context and common ground. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 105–108. Elsevier.

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*.

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 *[Proceedings] 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 *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.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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.

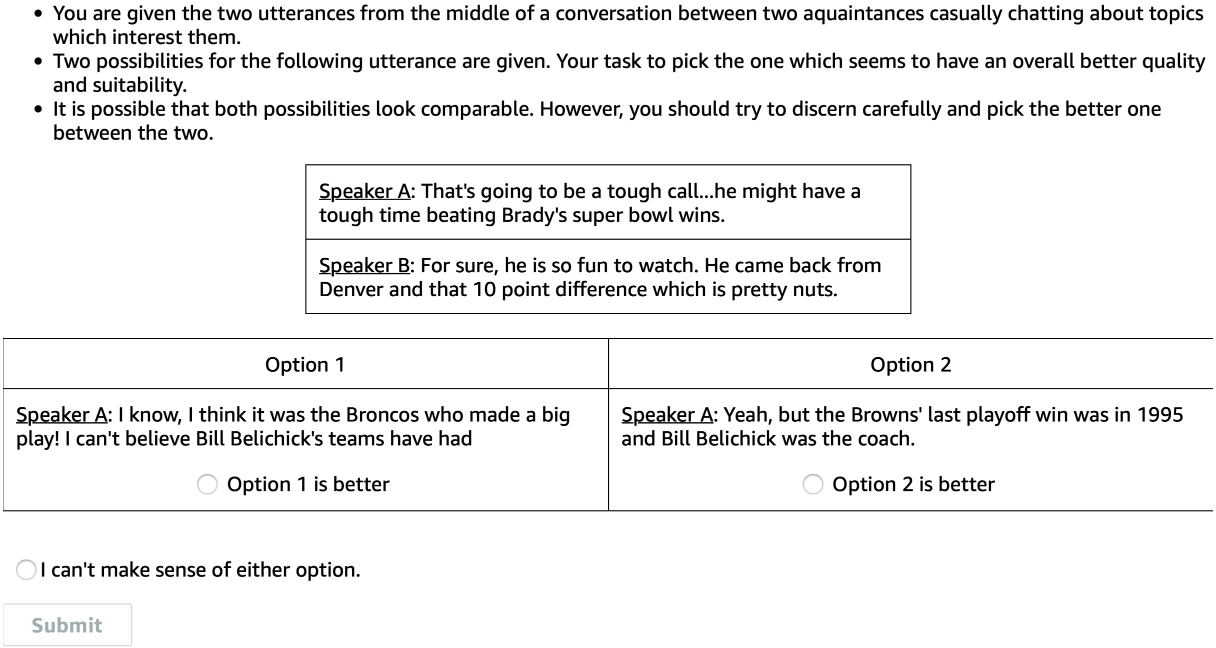
Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

A Appendix

A.1 Experimental 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 Annotation Interfaces



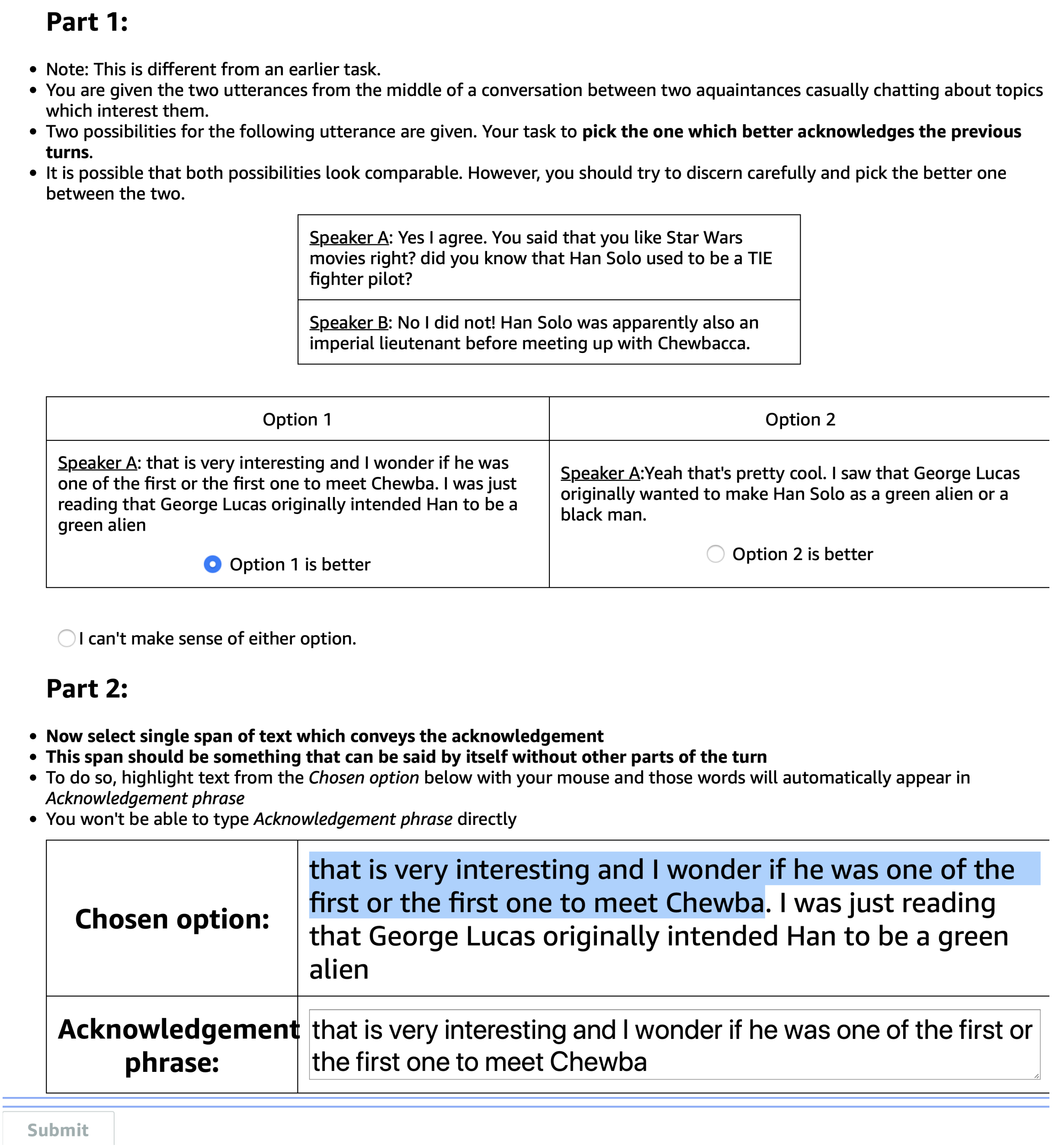
- You are given the two utterances from the middle of a conversation between two aquaintances casually chatting about topics which interest them.
- Two possibilities for the following utterance are given. Your task to pick the one which seems to have an overall better quality and suitability.
- It is possible that both possibilities look comparable. However, you should try to discern carefully and pick the better one between the two.

Speaker A: That's going to be a tough call...he might have a tough time beating Brady's super bowl wins.
Speaker B: For sure, he is so fun to watch. He came back from Denver and that 10 point difference which is pretty nuts.

Option 1	Option 2
Speaker A: I know, I think it was the Broncos who made a big play! I can't believe Bill Belichick's teams have had ○ Option 1 is better	Speaker A: Yeah, but the Browns' last playoff win was in 1995 and Bill Belichick was the coach. ○ Option 2 is better

○ I can't make sense of either option.

Figure 6: Annotation interface for Best PMI v/s rest



Part 1:

- Note: This is different from an earlier task.
- You are given the two utterances from the middle of a conversation between two aquaintances casually chatting about topics which interest them.
- Two possibilities for the following utterance are given. Your task to **pick the one which better acknowledges the previous turns**.
- It is possible that both possibilities look comparable. However, you should try to discern carefully and pick the better one between the two.

Speaker A: Yes I agree. You said that you like Star Wars movies right? did you know that Han Solo used to be a TIE fighter pilot?
Speaker B: No I did not! Han Solo was apparently also an imperial lieutenant before meeting up with Chewbacca.

Option 1	Option 2
Speaker A: 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 ● Option 1 is better	Speaker A:Yeah that's pretty cool. I saw that George Lucas originally wanted to make Han Solo as a green alien or a black man. ○ Option 2 is better

○ I can't make sense of either option.

Part 2:

- **Now select single span of text which conveys the acknowledgement**
- **This span should be something that can be said by itself without other parts of the turn**
- To do so, highlight text from the *Chosen option* below with your mouse and those words will automatically appear in *Acknowledgement phrase*
- You won't be able to type *Acknowledgement phrase* directly

Chosen option:	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
Acknowledgement phrase:	that is very interesting and I wonder if he was one of the first or the first one to meet Chewba

Figure 7: Annotation interface for acknowledgement differences due to pcmi_h