Learning an Unreferenced Metric for Online Dialogue Evaluation

Koustuv Sinha,[*,1,2,3] **Prasanna Parthasarathi,** [1,2] **Jasmine Wang,** [1]
Ryan Lowe, [1,2,4] **William L. Hamilton,** [1,2] and **Joelle Pineau** [1,2,3]

[1] School of Computer Science, McGill University, Canada
[2] Quebec Artificial Intelligence Institute (Mila), Canada
[3] Facebook AI Research (FAIR), Montreal, Canada
[4] OpenAI

Abstract

Evaluating the quality of a dialogue interaction between two agents is a difficult task, especially in open-domain chit-chat style dialogue. There have been recent efforts to develop automatic dialogue evaluation metrics, but most of them do not generalize to unseen datasets and/or need a human-generated reference response during inference, making it infeasible for online evaluation. Here, we propose an *unreferenced* automated evaluation metric that uses large pre-trained language models to extract latent representations of utterances, and leverages the temporal transitions that exist between them. We show that our model achieves higher correlation with human annotations in an online setting, while not requiring true responses for comparison during inference.

1 Introduction

Recent approaches in deep neural language generation have opened new possibilities in dialogue generation (Serban et al., 2017; Weston et al., 2018). Most of the current language generation efforts are centered around language modelling or machine translation (Ott et al., 2018), which are evaluated by comparing directly against the reference sentences. In dialogue, however, comparing with a single reference response is difficult, as there can be many reasonable responses given a context that have nothing to do with each other (Liu et al., 2016). Still, dialogue research papers tend to report scores based on *word-overlap* metrics from the machine translation literature (e.g. BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014)). However word-overlap metrics aggressively penalize the generated response based on lexical differences with the ground truth and correlate poorly to human judgements (Liu et al., 2016).
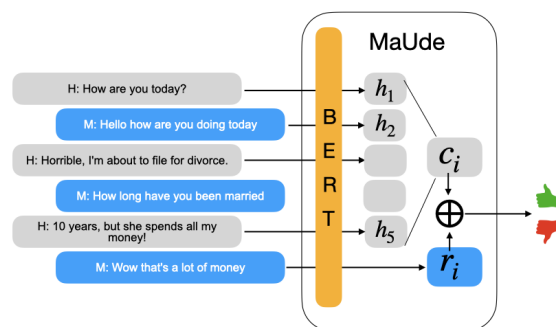

Figure 1: Model architecture for MaUdE, which is an unsupervised unreferenced metric for dialog evaluation.

One can build dialogue evaluation metrics in two ways: *referenced* metrics, which compare the generated response with a provided ground-truth response (such as the above word-overlap metrics), or an *unreferenced* metrics, which evaluate the generated response without any such comparison. Lowe et al. (2017) propose a *learned referenced* metric named ADEM, which learns an alignment score between context and response to predict human score annotations. However, since the score is trained to mimic human judgements, it requires collecting large-scale human annotations on the dataset in question and cannot be easily applicable to new datasets (Lowe, 2019).

Recently, Tao et al. (2017) proposed a hybrid referenced-unreferenced metric named RUBER, where the metric is trained without requiring human responses by bootstrapping negative samples directly from the dataset. However, referenced metrics (including RUBER, as it is part referenced) are not feasible for evaluation of dialogue models in an *online* setting—when the model is pitched against a human agent (model-human) or a model agent (model-model)—due to lack of a reference response. In this setting, models are usually eval-

uated directly by humans, which is costly and requires careful annotator training (Li et al., 2019).

The contributions of this paper are (1) a completely unsupervised unreferenced metric MAUDE (**M**etric for **a**utomatic **U**nreferenced **d**ialogue **e**valuation), which leverages state-of-the-art pretrained language models (Devlin et al., 2018; Sanh et al., 2019), combined with a novel discourse-structure aware text encoder and contrastive training approach; and (2) results showing that MAUDE has good correlation with human judgements.

2 Background

We consider the problem of evaluating the response of a dialogue system, where an agent is provided with a sequence of sentences (or utterances) $c = \{u_1, u_2, ..., u_n\}$ (termed as *context*) to generate a *response* $r = u_{n+1}$. Each utterance, u_i, can be represented as a set of words $u_i = \{w_1, w_2, ..., w_n\}$. An utterance u_i can be represented as a vector as $\mathbf{h}_i = f_e(u_i)$, where f_e is an encoder that encodes the words into a fixed vector representation.

This work focuses on the evaluation of *generative neural dialogue models*, which typically consist of an encoder-decoder style architecture that is trained to generate u_{n+1} word-by-word (Serban et al., 2017). The response of a generative model is typically evaluated by comparing with the ground-truth response using various automatic word-overlap metrics, such as BLEU or METEOR. These metrics, along with ADEM and RUBER, are essentially *single-step* evaluation metrics, where a score is calculated for each context-response pair. If a dialogue D_i contains n utterances, we can extract $n-1$ context-response pairs : $(c_1 : \{u_1\}, r_1 : \{u_2\}), (c_2 : \{u_1, u_2\}, r_2 : \{u_3\}), \ldots, (c_{n-1} : \{u_1 \ldots u_{n-1}\}, r_{n-1} : u_n)$. In this paper, we are interested in devising a scalar metric that can evaluate the quality of a context-response pair: $\text{score}(c_i, r_i) = \mathcal{R} \in (0, 1)$. A key benefit of this approach is that this metric can be used to evaluate online and also for better training and optimization, as it provides partial credit during response generation.

3 Proposed model

We propose a new model, MAUDE, for online unreferenced dialogue evaluation. We first describe the general framework behind MAUDE, which is inspired by the task of measuring alignment in natural language inference (NLI) (Williams et al., 2017). It involves training text encoders via noise contrastive estimation (NCE) to distinguish between valid dialogue responses and carefully generated negative examples. Following this, we introduce our novel text encoder that is designed to leverage the unique structural properties of dialogue.

MAUDE is designed to output a scalar $\text{score}(c_i, r_i) = \mathcal{R} \in (0, 1)$, which measures how appropriate a response r_i is given a dialogue context c_i. This task is analogous to measuring alignment in NLI, but instead of measuring entailment or contradiction, our notion of alignment aims to quantify the *quality* of a dialogue response. As in NLI, we approach this task by defining encoders $f_e^\theta(c)$ and $f_e^\theta(r)$ to encode the context and response, a combination function $f_{comb}(.)$ to combine the representations, and a final classifier $f_t(.)$, which outputs the alignment score:

$$\text{score}(c, r) = \sigma(f_t(f_{comb}(f_e^{\theta_1}(c), f_e^{\theta_2}(r)))). \quad (1)$$

The key idea behind an unreferenced dialogue metric is the use of Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010) for training. Specifically, we train the model to differentiate between a correct response ($\text{score}(c, r) \to 1$), and a negative response ($\text{score}(c, \hat{r}) \to 0$), where $\hat{r}$ represents a candidate false response for the given context c. The loss to minimize contains one positive example and a range of negative examples chosen from a sampling policy $P(\hat{r})$:

$$\mathcal{L} = -\log(\text{score}(c, r)) - \mathbb{E}_{\hat{r} \sim P(\hat{r})} \log(-\text{score}(c, \hat{r})). \quad (2)$$

The sampling policy $P(\hat{r})$ consists of *syntactic* and *semantic* negative samples.

Syntactic negative samples. We consider three variants of syntax level adversarial samples: *word-order* (shuffling the ordering of the words of r), *word-drop* (dropping $x\%$ of words in r) and *word-repeat* (randomly repeating words in r).

Semantic negative samples. We also consider three variants of negative samples that are syntactically well formed, but represent corruption in the semantic space. First, we choose a response r_j which is chosen at random from a different dialogue such that $r_j \neq r_i$ (*random utterance*). Second, we use a pre-trained seq2seq model on the dataset, and pair random seq2seq generated response with r_i (*random seq2seq*). Third, to provide a bigger variation of semantically negative samples, for each r_i we generate high-quality paraphrases

r_i^b using Back-Translation (Edunov et al., 2018). We pair random Back-Translations r_j^b with r_i as in the above setup (*random back-translation*). We also provide the paired r_i^b as positive example for the models to learn variation in semantic similarity. We further discuss the effect of different sampling policies in Appendix C.

Dialogue-structure aware encoder. Traditional NLI approaches (e.g., Conneau et al. (2017)) use the general setup of Equation 1 to score context-response pairs. The encoder f_e is typically a Bidirectional LSTM—or, more recently, a BERT-based model (Devlin et al., 2018), which uses a large pre-trained language model. f_{comb} is defined as in Conneau et al. (2017):

$$f_{comb}(u, v) = \text{concat}([u, v, u * v, u - v]). \quad (3)$$

However, the standard text encoders used in these traditional NLI approaches ignore the temporal structure of dialogues, which is critical in our setting where the context is composed of a sequence of distinct utterances, with natural and stereotypical transitions between them. (See Appendix A for a qualitative analysis of these transitions). Thus we propose a specialized text encoder for MAUDE, which uses a BERT-based encoder f_e^{BERT} but additionally models dialogue transitions using a recurrent neural network:

$$\mathbf{h}_{u_i} = \mathbf{D}_g f_e^{\text{BERT}}(u_i),$$
$$\mathbf{h}'_{u_{i+1}} = f_R(\mathbf{h}_{u_i}, \mathbf{h}'_{u_i}),$$
$$\mathbf{c}_i = \mathbf{W}.\text{pool}_{\forall t \in \{u_1, ..., u_{n-1}\}}(\mathbf{h}'_t)$$
$$\text{score}(c_i, r_i) = \sigma(f_t([\mathbf{h}_{r_i}, \mathbf{c}_i, \mathbf{h}_{r_i} * \mathbf{c}_i, \mathbf{h}_{r_i} - \mathbf{c}_i])), \quad (4)$$

where $\mathbf{h}_{u_i} \in \mathcal{R}^d$ is a downsampled BERT representation of the utterance u_i (using a global learned mapping $\mathbf{D}_g \in \mathcal{R}^{B \times d}$). $\mathbf{h}'_{u_i}$ is the hidden representation of f_R for u_i, where f_R is a Bidirectional LSTM. The final representation of the dialogue context is learned by pooling the individual hidden states of the RNN using max-pool (Equation 4). This context representation is mapped into the response vector space using weight $\mathbf{W}$, to obtain $\mathbf{c}_i$. We then learn the alignment score between the context $\mathbf{c}_i$ and response r_i's representation $\mathbf{h}_{r_i}$ following Equation 1, by using the combination function f_{comb} being the same as in Equation 3.

4 Experiments

To empirically evaluate our proposed unreferenced dialogue evaluation metric, we are interested in answering the following key research questions:

- Q1: How robust is our proposed metric on different types of responses?

- Q2: How well does the self-supervised metric correlate with human judgements?

Datasets. For training MAUDE, we use PersonaChat (Zhang et al., 2018), a large-scale open-domain chit-chat style dataset which is collected by human-human conversations over provided *user persona*. We extract and process the dataset using ParlAI (Miller et al.) platform. We use the public train split for our training and validation, and the public validation split for testing. We use the human-human and human-model data collected by See et al. (2019) for correlation analysis, where the models themselves are trained on PersonaChat.

Baselines. We use InferSent (Conneau et al., 2017) and unreferenced RUBER as LSTM-based baselines. We also compare against BERT-NLI, which is the same as the InferSent model but with the LSTM encoder replaced with a pre-trained BERT encoder. Note that these baselines can be viewed as ablations of the MAUDE framework using simplified text encoders, since we use the same NCE training loss to provide a fair comparison. Also, note that in practice, we use DistilBERT (Sanh et al., 2019) instead of BERT in both MAUDE and the BERT-NLI baseline (and thus we refer to the BERT-NLI baseline as DistilBERT-NLI).[1]

4.1 Evaluating MAUDE on different types of responses

We first analyze the robustness of MAUDE by comparing with the baselines, by using the same NCE training for *all the models* for fairness. We evaluate the models on the *difference* score, $\Delta = \text{score}(c, r_{\text{ground-truth}}) - \text{score}(c, r)$ (Table 6). Δ provides an insight on the range of score function. An optimal metric would cover the full range of good and bad responses. We evaluate response r in three settings: *Semantic Positive*: responses that are semantically equivalent to the ground truth response; *Semantic Negative*: responses that are semantically opposite to the ground truth response; and *Syntactic*

---

[1]DistilBERT is the same BERT encoder with significantly reduced memory footprint and training time, which is trained by knowledge distillation (Bucilu et al., 2006; Hinton et al., 2015) on the large pre-trained model of BERT.

		R	IS	DNLI	M
Semantic Positive ↓	BackTranslation	0.249	0.278	**0.024**	0.070
	Seq2Seq	0.342	0.362	**0.174**	0.308
Semantic Negative ↑	Random Utterance	0.152	0.209	0.147	**0.287**
	Random Seq2Seq	0.402	0.435	0.344	**0.585**
Syntactic Negative ↑	Word Drop	0.342	**0.367**	0.261	0.3
	Word Order	0.392	0.409	0.671	**0.726**
	Word Repeat	0.432	0.461	0.782	**0.872**

Table 1: Metric score evaluation ($\Delta = \text{score}(c, r_{\text{ground-truth}}) - \text{score}(c, r)$) between RUBER (R), InferSent (IS), DistilBERT-NLI (DNI) and MAUDE (M) on PersonaChat dataset's public validation set. For Semantic Positive tests, lower Δ is better; for all Negative tests higher Δ is better.

	R	IS	DNLI	M
Fluency	0.322	0.246	**0.443**	0.37
Engagingness	0.204	0.091	0.192	**0.232**
Humanness	0.057	-0.108	**0.129**	0.095
Making Sense	0.0	0.005	**0.256**	0.208
Inquisitiveness	0.583	0.589	0.598	**0.728**
Interestingness	**0.275**	0.119	0.135	0.24
Avoiding Repetition	**0.093**	-0.118	-0.039	-0.035
Listening	0.061	-0.086	**0.124**	0.112
Mean	0.199	0.092	0.23	**0.244**

Table 2: Correlation with calibrated scores between RUBER (R), InferSent (IS), DistilBERT-NLI (DNI) and MAUDE (M) when trained on PersonaChat dataset

Negative: responses that have been adversarially modified in the lexical units. Ideally, we would want $\Delta \to 1$ for semantic and syntactic negative responses, $\Delta \to 0$ for semantic positive responses.

We observe that the MAUDE scores perform robustly across all the setups. RUBER and InferSent baselines are weak, quite understandably so because they cannot leverage the large pre-trained language model data and thus is poor at generalization. DistilBERT-NLI baseline performs significantly better than InferSent and RUBER, while MAUDE scores even better and more consistently overall. We provide a detailed ablation of various training scenarios as well as the absolute raw Δ scores in Appendix C. We also observe both MAUDE and DistilBERT-NLI to be more robust on zero-shot generalization to different datasets, the results of which are available in Appendix B.

4.2 Correlation with human judgements

Metrics are evaluated on correlation with human judgements (Lowe et al., 2017; Tao et al., 2017), or by evaluating the responses of a generative model trained on the metric (Wieting et al., 2019), by human evaluation. However, this introduces a bias either during the questionnaire setup or during data post-processing in favor of the proposed metric. In this work, we refrain from collecting human annotations ourselves, but refer to the recent work by See et al. (2019) on PersonaChat dataset. Thus, the evaluation of our metric is less subject to bias.

See et al. (2019) conducted a large-scale human evaluation of 28 model configurations to study the effect of controllable attributes in dialogue generation. We use the publicly released model-human and human-human chat logs from See et al. (2019) to generate the scores on our models, and correlate them with the associated human judgement on a Likert scale. See et al. (2019) propose to use a *multi-step* evaluation methodology, where the hu-

man annotators rate the entire dialogue and not a context-response pair. On the other hand, our setup is essentially a *single-step* evaluation method. To align our scores with the multi-turn evaluation, we average the individual turns to get an aggregate score for a given dialogue.
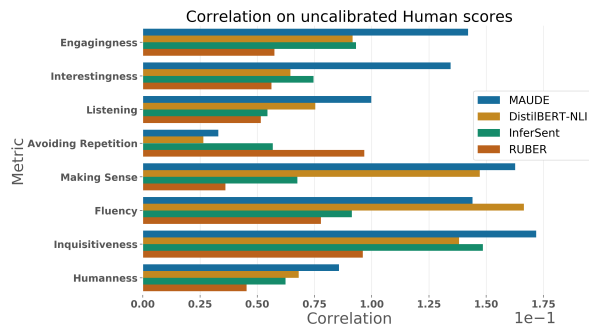

Figure 2: Human correlation on un-calibrated scores collected on PersonaChat dataset (Zhang et al., 2018), for MAUDE, DistilBERT-NLI, InferSent and RUBER

We investigate the correlation between the scores and uncalibrated individual human scores from 100 crowdworkers (Fig. 2), as well as aggregated scores released by See et al. (2019) which are adjusted for annotator variance by using Bayesian calibration (Kulikov et al., 2018) (Table 2). In all cases, we report Spearman's correlation coefficients.

For uncalibrated human judgements, we observe MAUDE having higher relative correlation in 6 out of 8 quality measures. Interestingly, in case of calibrated human judgements, DistilBERT proves to be better in half of the quality measures. MAUDE achieves marginally better overall correlation for calibrated human judgements, due to significantly strong correlation on specifically two measures: Interestingness and Engagingness. These measures answers the questions *"How interesting or boring did you find this conversation?"* and *"How much did you enjoy talking to this user?"*. (Refer to Appendix B of See et al. (2019) for the full

list of questions). Overall, using large pre-trained language models provides significant boost in the human correlation scores.

5 Conclusion

In this work, we explore the feasibility of learning an automatic dialogue evaluation metric by leveraging pre-trained language models and the temporal structure of dialogue. We propose MAUDE, which is an unreferenced dialogue evaluation metric that leverages sentence representations from large pre-trained language models, and is trained via Noise Contrastive Estimation. MAUDE also learns a recurrent neural network to model the transition between the utterances in a dialogue, allowing it to correlate better with human annotations. This is a good indication that MAUDE can be used to evaluate online dialogue conversations. Since it provides immediate continuous rewards and at the single-step level, MAUDE can be also be used to optimize and train better dialogue generation models, which we want to pursue as future work.

Acknowledgements

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. *arXiv*.

Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP*. MultiWoz CORPUS licensed under CC-BY 4.0.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of ACL*.

W.A. Falcon. 2019. Pytorch lightning. https://github.com/williamFalcon/pytorch-lightning.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv*.

Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *arXiv*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of IJCNLP*.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. *arXiv*.

Ryan Lowe. 2019. A Retrospective for "Towards an Automatic Turing Test - Learning to Evaluate Dialogue Responses". *ML Retrospectives*.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. *arXiv*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. Parlai: A dialog research software platform. *arXiv*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. *arXiv*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of AAAI*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. *arXiv*.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:Training Neural Machine Translation with Semantic Similarity. In *Proceedings of ACL*, Florence, Italy.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv*.

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? *arXiv*.

A Temporal Structure

We hypothesize that a *good* encoding function can capture the structure that exists in dialogue. Often this translates to capturing the semantics, coherency in dialogue which are some of the key attributes of a conversation. Formally, we propose using a function $f_t^{D_i}$ which maps one utterance to the next.

$$\mathbf{h}_{u_{i+1}} = f_t^{D_i}(\mathbf{h}_{u_i}) \qquad (5)$$

To define a *good* encoding function, we turn to pre-trained language models. These models are typically trained on large corpus and achieve state-of-the-art results on a range of language understanding tasks (Ott et al., 2018). To validate our hypothesis, we use a pre-trained (and fine-tuned) BERT (Devlin et al., 2018) as f_e. We compute $h_{u_i} = f_e(u_i) \forall u_i \in D$, and learn a linear classifier to predict an approximate position of the $u_i \in D_i$. The task has details in its design, in the case of goal-oriented dialogues the vocabulary differs in different parts of the conversation and in chitchat dialogues it cannot be said. To experiment, we choose PersonaChat (Zhang et al., 2018) and DailyDialog (Li et al., 2017) to be nominal of chit-chat style data, and Frames (Asri et al., 2017) and Multi-WOZ (Budzianowski et al., 2018) for goal-oriented data.

We encode every consecutive pairs of the utterances with a % score, t, that denotes its occurrence after the completion of $t\%$ of dialogue.

$$t_{u_p} = \frac{\text{index}_{u_p} + 1}{k} \qquad (6)$$

where index_{u_p} denote the average of the indices in the pair of the utterances and k denote the total number of utterances in dialogue.

Now, we pre-define the number of bins *B*. We split the range 0-100 into *B* non-overlapping sets(every set will have min and max denoted by s_{min}^i and s_{max}^i respectively). We parse every dialogue in the dataset, and place the encoding of every utterance pair in the corresponding bin.

$$bin_{u_p} = \{i \mid t_{u_p} > s_{min}^i \& s_{max}^i > t_{u_p}\} \qquad (7)$$

We then use Linear Discriminant Analysis (LDA) to predict the bin of each utterance u_i in the dialogue after converting the high dimensional embedding into 2 dimensions. LDA provides the best possible class conditioned representation of data. This gives us a downsampled representation of each utterance u_i which we plot as shown in Figure 3. The reduction on BERT encoding to 2-dimensions shows that BERT is useful in nudging the encoded utterances towards useful structures. We see well defined clusters in goal-oriented but not-so-well-defined clusters in open domain dialogues. This is reasonable to expect and intuitive.

B Generalization on unseen dialog datasets

In order for a dialogue evaluation metric to be useful, one has to evaluate how it generalizes to unseen data. We performed the evaluation using our trained models on PersonaChat dataset, and then evaluated them *zero-shot* on two goal-oriented datasets, Frames (Asri et al., 2017) and MultiWoz (Budzianowski et al., 2018), and one chit-chat style dataset: Daily Dialog (Li et al., 2017) (Table 3). We find BERT-based models are significantly better at generalization than InferSent or RUBER, with MAUDE marginally better than DistilBERT-NLI baseline. MAUDE has the biggest impact on generalization to DailyDialog dataset, which suggests that it captures the commonalities of chit-chat style dialogue from PersonaChat. Surprisingly, generalization gets significantly better of BERT-based models on goal-oriented datasets as well. This suggests that irrespective of the nature of dialogue, pre-training helps because it contains the information common to English language lexical items.

C Noise Contrastive Estimation training ablations

The choice of negative samples (Section 3) for Noise Contrastive Estimation can have a large impact on the test-time scores of the metrics. In this section, we show the effect when we train only using syntactic negative samples (Table 4) and only semantic negative samples (Table 5). For comparison, we show the full results when trained using both of the sampling scheme in Table 6. We find overall training only using either syntactic or semantic negative samples achieve less Δ than training using both of the schemes. All models achieve high scores on the semantic positive samples when only trained with syntactical adversaries. However, training only with syntactical negative samples results in adverse effect on detecting semantic negative items.

Datasets		DailyDialog		Frames		MultiWOZ	
Model	Eval Mode	Score	Δ	Score	Δ	Score	Δ
RUBER	+	0.173 ± 0.168		0.211 ± 0.172		0.253 ± 0.177	
	−	0.063 ± 0.092	0.11	0.102 ± 0.114	0.109	0.121 ± 0.123	0.123
InferSent	+	0.163 ± 0.184		0.215 ± 0.186		0.277 ± 0.200	
	−	0.050 ± 0.085	0.113	0.109 ± 0.128	0.106	0.127 ± 0.133	0.15
DistilBERT NLI	+	0.885 ± 0.166		0.744 ± 0.203		0.840 ± 0.189	
	−	0.575 ± 0.316	0.31	0.538 ± 0.330	0.206	0.566 ± 0.333	0.274
MAUDE	+	0.782 ± 0.248		0.661 ± 0.293		0.758 ± 0.265	
	−	0.431 ± 0.300	**0.351**	0.454 ± 0.358	**0.207**	0.483 ± 0.345	**0.275**

Table 3: Zero-shot generalization results on DailyDialog, Frames and MultiWOZ dataset for the baselines and MAUDE. + denotes semantic positive responses, and − denotes semantic negative responses.

PersonaChat Dataset	Model	RUBER		InferSent		DistilBERT NLI		MAUDE	
	Training Modes	Only Semantics		Only Semantics		Only Semantics		Only Semantics	
	Evaluation Modes	Score	Δ	Score	Δ	Score	Δ	Score	Δ
Semantic Positive	Gold Truth Response	$0.443_{\pm0.197}$	0	$0.466_{\pm0.215}$	0	$0.746_{\pm0.236}$	0	**0.789**$_{\pm0.244}$	0
	BackTranslation	$0.296_{\pm0.198}$	0.147	$0.273_{\pm0.195}$	0.192	**0.766**$_{\pm0.235}$	-0.02	$0.723_{\pm0.277}$	**0.066**
	Seq2Seq	$0.082_{\pm0.163}$	0.361	$0.10_{\pm0.184}$	0.367	**0.46**$_{\pm0.357}$	**0.286**	$0.428_{\pm0.390}$	0.361
Semantic Negative	Random Utterance	$0.299_{\pm0.203}$	0.144	**0.287**$_{\pm0.208}$	0.178	$0.489_{\pm0.306}$	0.257	$0.388_{\pm0.335}$	**0.40**
	Random Seq2Seq	**0.028**$_{\pm0.077}$	0.415	$0.036_{\pm0.082}$	0.429	$0.237_{\pm0.283}$	0.529	$0.16_{\pm0.26}$	**0.629**
Syntactic Negative	Word Drop	$0.334_{\pm0.206}$	0.109	**0.308**$_{\pm0.217}$	**0.158**	$0.802_{\pm0.224}$	-0.056	$0.73_{\pm0.29}$	0.059
	Word Order	**0.472**$_{\pm0.169}$	-0.029	$0.482_{\pm0.19}$	-0.016	$0.685_{\pm0.284}$	0.061	$0.58_{\pm0.35}$	**0.209**
	Word Repeat	$0.255_{\pm0.24}$	0.188	**0.153**$_{\pm0.198}$	0.312	$0.657_{\pm0.331}$	0.089	$0.44_{\pm0.39}$	**0.349**

Table 4: Metric score evaluation between InferSent, DistilBERT-NLI and MAUDE on PersonaChat dataset, trained on $P(\hat{r})$ = Semantics. Bold scores represent the best individual scores, and bold with blue represents the best difference with the true response.

PersonaChat Dataset	Model	RUBER		InferSent		DistilBERT NLI		MAUDE	
	Training Modes	Only Syntax		Only Syntax		Only Syntax		Only Syntax	
	Evaluation Modes	Score	Δ	Score	Δ	Score	Δ	Score	Δ
Semantic Positive	Gold Truth Response	$0.891_{\pm0.225}$	0	$0.893_{\pm0.231}$	0	$0.986_{\pm0.088}$	0	**0.99**$_{\pm0.07}$	0
	BackTranslation	$0.687_{\pm0.363}$	0.204	$0.672_{\pm0.387}$	0.221	$0.877_{\pm0.268}$	0.109	**0.91**$_{\pm0.23}$	**0.08**
	Seq2Seq	$0.929_{\pm0.187}$	-0.038	$0.949_{\pm0.146}$	-0.055	$0.996_{\pm0.048}$	-0.01	**0.99**$_{\pm0.05}$	**0.00**
Semantic Negative	Random Utterance	$0.869_{\pm0.248}$	0.022	**0.835**$_{\pm0.294}$	**0.058**	$0.977_{\pm0.116}$	0.009	$0.97_{\pm0.13}$	0.02
	Random Seq2Seq	$0.915_{\pm0.196}$	-0.024	**0.904**$_{\pm0.206}$	-0.011	$0.994_{\pm0.057}$	-0.008	$0.99_{\pm0.08}$	**0**
Syntactic Negative	Word Drop	$0.119_{\pm0.255}$	0.772	**0.105**$_{\pm0.243}$	**0.788**	$0.373_{\pm0.414}$	0.613	$0.41_{\pm0.44}$	0.584
	Word Order	$0.021_{\pm0.101}$	0.87	**0.015**$_{\pm0.0915}$	0.878	$0.064_{\pm0.194}$	0.922	$0.07_{\pm0.21}$	**0.928**
	Word Repeat	**0.001**$_{\pm0.007}$	0.89	**0.001**$_{\pm0.020}$	0.893	$0.006_{\pm0.057}$	0.980	$0.01_{\pm0.06}$	**0.981**

Table 5: Metric score evaluation between InferSent, DistilBERT-NLI and MAUDE on PersonaChat dataset, trained on $P(\hat{r})$ = Syntax. Bold scores represent the best individual scores, and bold with blue represents the best difference with the true response.
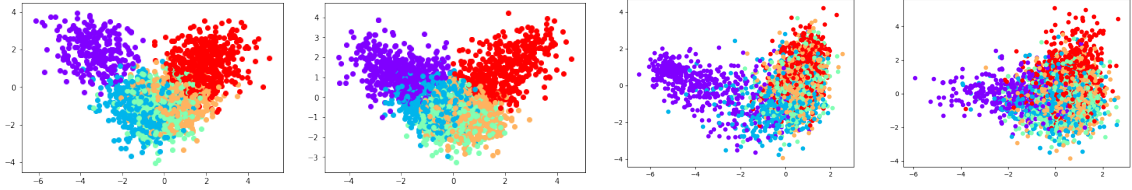
Figure 3: From left to right, LDA downsampled representation of BERT on Frames (Goal oriented), MultiWOZ (Goal oriented), PersonaChat (chit-chat) and DailyDialog (chit-chat)

PersonaChat Dataset	Model	RUBER		InferSent		DistilBERT NLI		MAUDE	
	Training Modes	All		All		All		All	
	Evaluation Modes	Score	Δ	Score	Δ	Score	Δ	Score	Δ
	Gold Truth Response	$0.432_{\pm0.213}$	0	$0.462_{\pm0.254}$	0	$0.824_{\pm0.154}$	0	$\mathbf{0.909}_{\pm0.152}$	0
Semantic Positive	BackTranslation	$0.183_{\pm0.198}$	0.249	$0.184_{\pm0.218}$	0.278	$0.8_{\pm0.19}$	**0.024**	$\mathbf{0.838}_{\pm0.227}$	0.070
	Seq2Seq	$0.09_{\pm0.17}$	0.342	$0.10_{\pm0.184}$	0.362	$\mathbf{0.65}_{\pm0.287}$	**0.174**	$0.6008_{\pm0.38}$	0.308
Semantic Negative	Random Utterance	$0.28_{\pm0.21}$	0.152	$\mathbf{0.252}_{\pm0.236}$	0.209	$0.677_{\pm0.255}$	0.147	$0.621_{\pm0.344}$	**0.287**
	Random Seq2Seq	$0.03_{\pm0.09}$	0.402	$\mathbf{0.026}_{\pm0.079}$	0.435	$0.48_{\pm0.313}$	0.344	$0.323_{\pm0.355}$	**0.585**
Syntactic Negative	Word Drop	$\mathbf{0.09}_{\pm0.16}$	0.342	$0.094_{\pm0.17}$	**0.367**	$0.563_{\pm0.377}$	0.261	$0.609_{\pm0.401}$	0.3
	Word Order	$\mathbf{0.04}_{\pm0.10}$	0.392	$0.052_{\pm0.112}$	0.409	$0.153_{\pm0.29}$	0.671	$0.182_{\pm0.327}$	**0.726**
	Word Repeat	$\mathbf{0.00}_{\pm0.01}$	0.432	$0.001_{\pm0.010}$	0.461	$0.041_{\pm0.153}$	0.782	$0.036_{\pm0.151}$	**0.872**

Table 6: Metric score evaluation between InferSent, DistilBERT-NLI and MAUDE on PersonaChat dataset, trained on $P(\hat{r})$ = Syntax + Semantics. Bold scores represent the best individual scores, and bold with blue represents the best difference with the true response.

D Qualitative Evaluation

We investigate qualitatively how the scores of different models are on the online evaluation setup on See et al. (2019)'c collected data. In Figure 4, we show a sample conversation where a human evaluator is pitched against a strong model. Here, MAUDE scores correlate strongly with raw likert scores on different metrics. We observe that RUBER and InferSent baselines overall correlate negatively with the response. In Figure 5, we show another sample where a human evaluator is pitched against a weak model, which exhibits degenerate responses. We see both MAUDE and DistilBERT-NLI correlate strongly with human annotation and provides a very low score, compared to RUBER or InferSent.

Since we essentially cherry-picked good results, its only fair to show a similarly cherry-picked negative example of MAUDE. We sampled from responses where MAUDE scores are negatively correlated with human annotations on Inquisitiveness metric (5% of cases), and we show one of those responses in Figure 6. We notice how both DistilBERT-NLI and MAUDE fails to recognize the duplication of utterances which leads to a low overall score. This suggests there still exists room for improvement in developing MAUDE, possibly by training the model to detect degeneracy in the

context.

E Hyperparameters and Training Details

We performed rigorous hyperparameter search to tune our model MAUDE. We train MAUDE with downsampling, as we observe poor results when we run the recurrent network on top of 768 dimensions. Specifically, we downsample to 300 dimensions, which is the same used by our baselines RUBER and InferSent in their respective encoder representations. We also tested with the choice of either learning a PCA to downsample the BERT representations vs learning the mapping $\mathbf{D}_g$ (Equation 4), and found the latter producing better results. We keep the final decoder same for all models, which is a two layer MLP with hidden layer of size 200 dimensions and dropout 0.2. For BERT-based models (DistilBERT-NLI and MAUDE), we use HuggingFace Transformers (Wolf et al., 2019) to first fine-tune the training dataset on language model objective. We tested with training on frozen fine-tuned representations in our initial experiments, but fine-tuning end-to-end lead to better ablation scores. For all models we train using Adam optimizer with 0.0001 as the learning rate, early stopping till validation loss doesn't improve. For the sake of easy reproducibility, we use Pytorch Lightning (Falcon, 2019) framework. We used 8 Nvidia-TitanX GPUs

Figure 4: An example of dialogue conversation between human and a strong model, where MAUDE (M) score correlates positively with human annotations. Raw Likert scores for the entire dialogue are: Engagingness : 3, Interestingness : 3, Inquisitiveness : 2, Listening : 3, Avoiding Repetition : 3, Fluency : 4, Making Sense : 4, Humanness : 3, Persona retrieval : 1. Baselines are RUBER (R), InferSent (I) and BERT-NLI (B).

on a DGX Server Workstation to train faster using
Pytorch Distributed Data Parallel (DDP).

interesting_nidf_model_bfw_setting_08: Hi names jace and I drive a stickshift f150

Human: Hi, my name's Carroll and I drive a mini van. Haha

interesting_nidf_model_bfw_setting_08: Hondas kawasaki acura passa sx bu hace rrrr hui bien hao

R: 0.427   I: 0.210   B: 0.598   M: 0.036

Human: I'm sorry, what?

interesting_nidf_model_bfw_setting_08: Ll pinche peu pasa sya bla woh numa asi paino gila fer heaven's sake eh buh tsk dun wah arf whoo doo deh pon di kina dum dippy p

R: 0.573   I: 0.607   B: 0.000   M: 0.000

Human: I don't understand..

interesting_nidf_model_bfw_setting_08: Wher arre dey loke da fuc dah ni mach zu guan adrenalin howls ain ju ching wei oui kora maye zi goti ta creat vera

R: 0.638   I: 0.756   B: 0.075   M: 0.000

Human: What on earth are you talking about?

interesting_nidf_model_bfw_setting_08: Yo gotti isn'thing tu ruff boosie dontcha fet thos fishy beetles killa fien mus slurp the> mater ke maximo estas esta aking cayden tobe hur

R: 0.730   I: 0.6635   B: 0.002   M: 0.000

Human: Perhaps you ought to lie down.

interesting_nidf_model_bfw_setting_08: Ahahah como espanol es muy bueno jees loca xian acara ki kaya tai ko lelesfashionshop1 moe walpole kis $ions profesional geologist

R: 0.643   I: 0.535   B: 0.000   M: 0.000
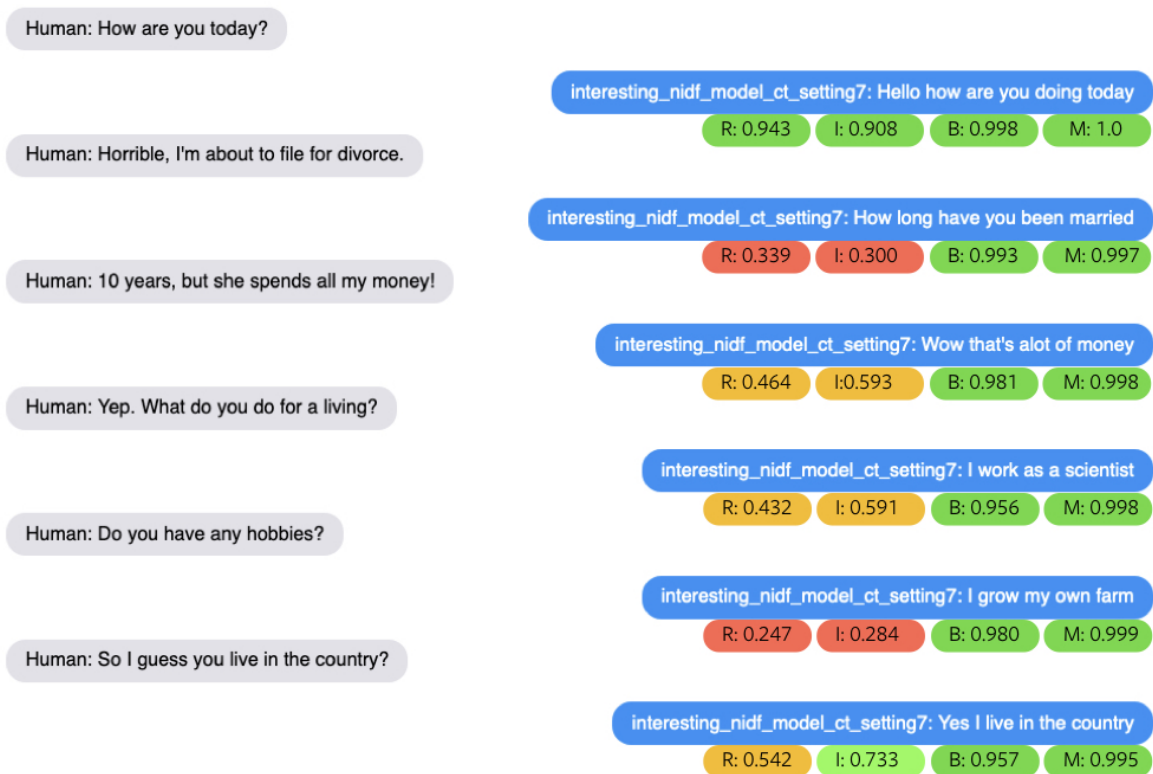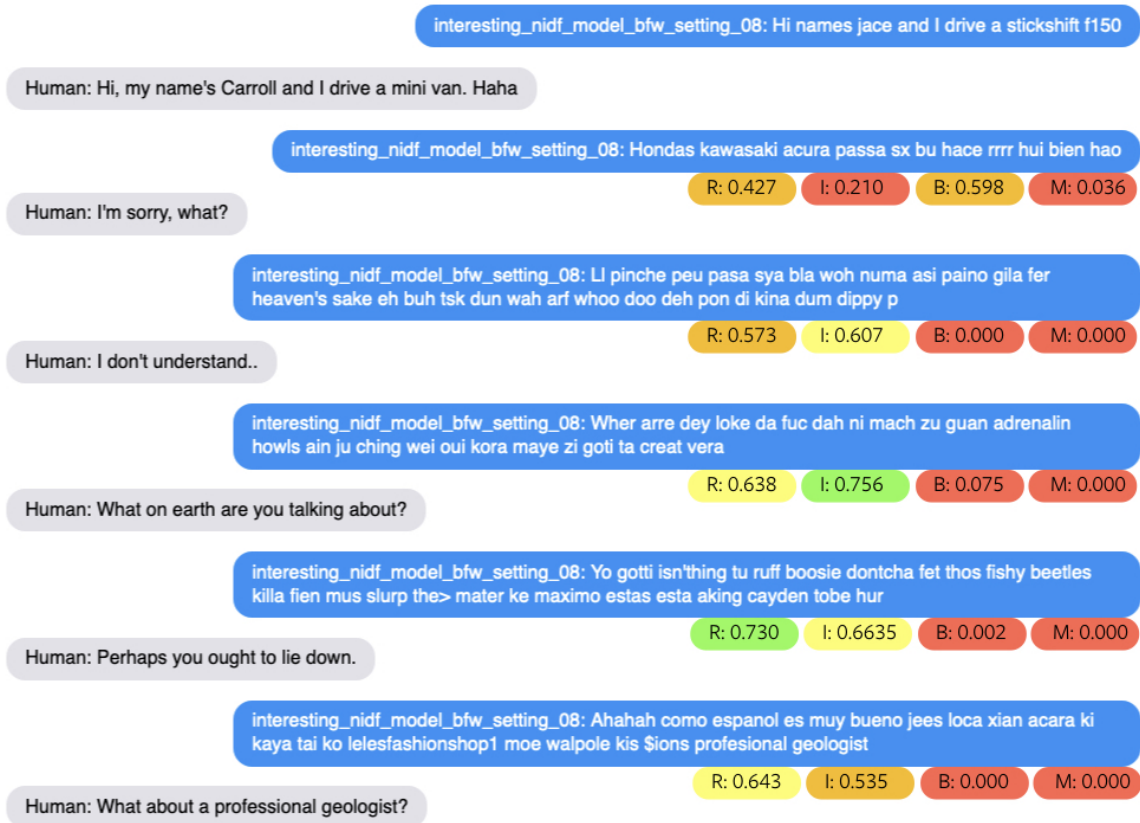
Human: What about a professional geologist?

Figure 5: An example of dialogue conversation between human and a weak model, where MAUDE (M) score correlates positively with human annotations. Raw Likert scores for the entire dialogue are: Engagingness : 1, Interestingness : 4, Inquisitiveness : 1, Listening : 1, Avoiding Repetition : 3, Fluency : 1, Making Sense : 2, Humanness : 1, Persona retrieval : 1. In our setup we only score responses *only* following a human response. Baselines are RUBER (R), InferSent (I) and BERT-NLI (B).
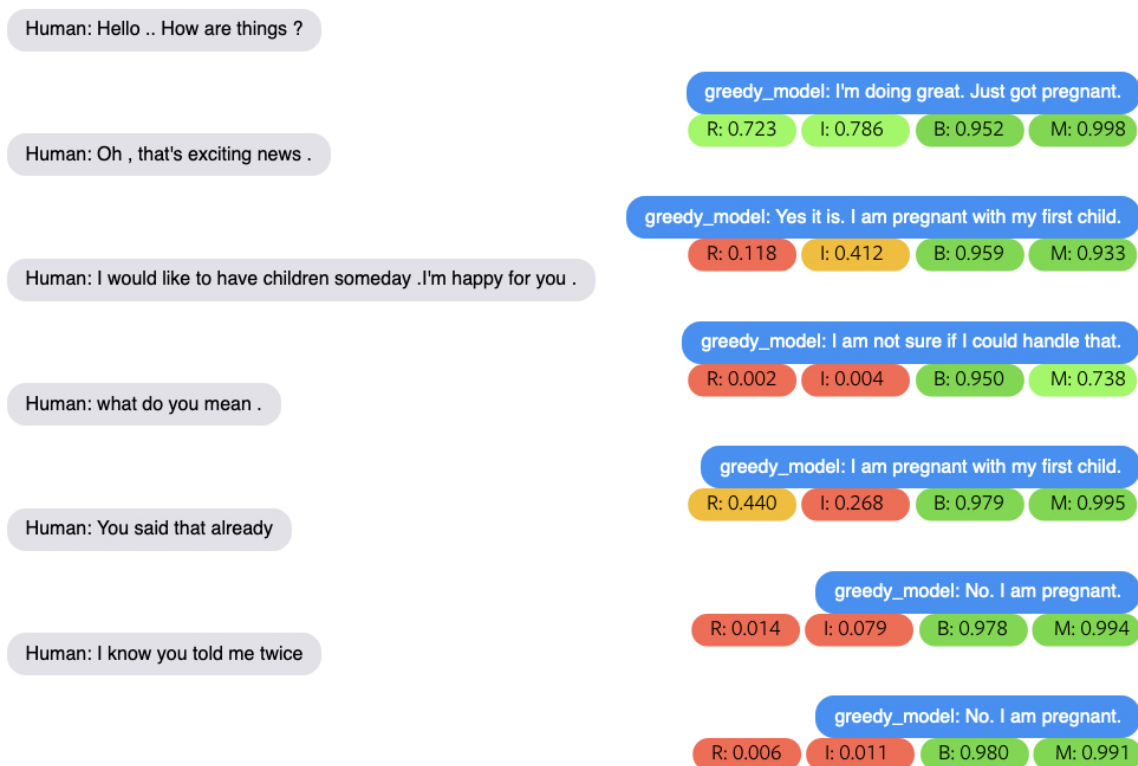
Figure 6: An example of dialogue conversation between human and a model, where MAUDE (M) score correlates negatively with human annotations. Raw Likert scores for the entire dialogue are: Engagingness : 1, Interestingness : 1, Inquisitiveness : 2, Listening : 2, Avoiding Repetition : 2, Fluency : 3, Making Sense : 4, Humanness : 2, Persona retrieval : 1. Baselines are RUBER (R), InferSent (I) and BERT-NLI (B).