Proxy Indicators for the Quality of Open-domain Dialogues

Rostislav Nedelchev
Smart Data Analytics Group, University of Bonn, Germany
rostislav.nedelchev@uni-bonn.de

Jens Lehmann
Smart Data Analytics Group,
University of Bonn, Germany
Fraunhofer IAIS,
Sankt Augustin and Dresden, Germany
jens.lehmann@cs.uni-bonn.de

Ricardo Usbeck
Semantic Systems Group,
University of Hamburg, Germany
Smart Data Analytics Group,
University of Bonn, Germany
ricardo.usbeck@uni-hamburg.de

Abstract

The automatic evaluation of open-domain dialogues remains a largely unsolved challenge. Thus, despite the abundance of work done in the field, human judges have to evaluate dialogues' quality. As a consequence, performing such evaluations at scale is usually expensive. This work investigates using a deep-learning model trained on the General Language Understanding Evaluation (GLUE) benchmark to serve as a quality indication of open-domain dialogues. The aim is to use the various GLUE tasks as different perspectives on judging the quality of conversation, thus reducing the need for additional training data or responses that serve as quality references. Due to this nature, the method can infer various quality metrics and derive a component-based overall score. We achieve statistically significant correlation coefficients of up to 0.7.

1 Introduction

Recently, dialogue systems powered by machine learning have gathered much attention from industry and academia alike (Chen et al., 2017). These systems have various applications, such as personal speech assistants, customer service, technical support, and training and education. In most cases, these systems are task-specific and help with tasks like booking a restaurant. Nevertheless, they can still benefit from open-domain conversational skills, such as the ability to chit-chat to enable natural dialogues, rather than repeating the input utterance like a parrot.

Nowadays, people working in this field have to use human annotators to evaluate the quality of a conversation (Dinan et al., 2019; Logacheva et al., 2018; Yoshino et al., 2019), which can be very costly in terms of resources. Thus, these systems' research and development could benefit significantly from an automated approach to evaluate conversations.

Research in the related fields of text summarization and machine translation has developed automated measures for evaluation. Some notable examples are, for the former, ROUGE (Lin, 2004) and, for the latter, BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). These are also adopted by works researching dialogue systems (Ritter et al., 2011; Serban et al., 2016; Yoshino et al., 2019). However, Liu et al. (2016) demonstrated that these metrics are not suitable for replacing human evaluators. Also, Sai et al. (2020) reported that even using multiple instead of single references and an overlap-based approach still underperforms. Thus, more advanced techniques are needed that consider the context and semantics of a dialogue.

Human annotators can distinguish bad from good quality dialogues from an intuitive perspective, not because they necessarily have been taught to do so. Instead, people have a notion of a fluent text or when a response is relevant (or not) to a previous utterance. Thus, this work's primary goal is to investigate a similar approach that mimics component-wise human intuition[1].

We investigate whether natural language processing (NLP) tasks can serve as proxy indicators for a conversation's quality. For that purpose, we use a fine-tuned BERT (Devlin et al., 2019) model trained on the GLUE benchmark (Wang et al., 2019). GLUE provides a comprehensive evaluation of general language understanding. We

[1]Resources to reproduce the work can be found at this link: https://github.com/SmartDataAnalytics/proxy_indicators

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7834–7855
November 7–11, 2021. ©2021 Association for Computational Linguistics

demonstrate that a few of the tasks exhibit a limited potential of serving as proxy indicators. The rest shows negative results.

2 Related Work

Lowe et al. (2017) propose a work that approximates human judgment using scored dialogues together with the context, reference response, and utterance generated by a dialogue system. However, the approach is hard to scale since reference responses and human annotation scores are still necessary. In another work, Tao et al. (2018) propose a method consisting of two parts. The first measures similarity to a reference response using a word embedding vector pooling. The second is a neural network that evaluates the relatedness of a reply given the context. The first component also uses reference responses, which are hard to acquire. Moreover, both approaches lack the interpretability of the scores they output regarding different dialogue quality features, such as coherency or fluency.

More recently, Ghandeharioun et al. (2019) propose a framework that uses self-play and two NLP tasks as an additional source of knowledge to evaluate dialogues in a multi-turn mode scenario. They perform an ablation study using sentiment and natural language inference as proxy supervision to see whether their system can better approximate human judgment. Their work shows that dialogue systems can benefit from using them. Also, Welleck et al. (2019) frame the dialogue consistency issue as a natural language inference problem and propose the DialogueNLI dataset. Its purpose is to benchmark a model's ability to select relevant utterances relative to a given context. Finally, Nedelchev et al. (2020b) offered to treat dialogue evaluation as an anomaly detection problem. Their results were negative and suggested that the approach suffers from insufficient training data.

Until very lately, Nedelchev et al. (2020a), Sai et al (2020), and Mehri et al. (2020) propose the usage of language models as indicator of quality. All of their approaches require no references or supervision. However, their proposed methods do not separate the different quality aspects and only indicate a dialogue's overall quality.

3 General Language Understanding Evaluation

This section briefly introduces the General Language Understanding Evaluation benchmark (Wang et al., 2019), its sub-tasks, and their relevance to this work. GLUE has two categories of tasks - single- and pairwise-sentence tasks. They provide annotated data for training models to solve various natural language understanding problems. The section also discusses how these NLP tasks could be related to dialogue evaluation since they are initially irrelevant to this paper's core topic. The presentation of each of the tasks follows:

3.1 Single-Sentence Tasks

Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2018) comprises samples in the English language that have scores for their grammatical correctness. Formally, this is a binary classification problem, where sentences are either acceptable (one) or unacceptable (zero) (Wang et al., 2019). To evaluate dialogues, CoLA can provide fluency measures that show how grammatically sound a conversation is.

Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) contains text excerpts from the movie reviews that have their sentiments annotated by humans as positive (one) or as negative (zero). Common sense would suggest that attitude provides no apparent relation to dialogue quality. Nonetheless, Ghandeharioun et al. (2019) perform an ablation study as part of their work to see if knowledge distillation based on sentiment offers any benefits to evaluating a conversation. Their research shows that there can be an improvement depending on the neural network model and the target dataset. So, we investigate how it relates to annotator scoring on dialogue evaluation.

3.2 Pairwise-Sentence Tasks

The pairwise-sentence tasks consider a pair of utterances that appear sequentially in a dialogue.

Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) is a dataset of sentence pairs extracted from news media, where each couple has scores as having the same meaning or not. Formally, it is a binary classification problem. A paraphrase has a label as

positive, and non-semantic equivalence is negative. In the context of dialogues, a machine learning prediction for this task could imply that a response to an utterance is just repeating the former. At the same time, a partial degree could be suggesting some relevance. The negative case does not have a straightforward interpretation.

Quora Question Pairs (QQP) [2] is a corpus of question pairs extracted from the community question-answering platform Quora. Similar to MRPC, The focus is to flag a duo of questions as having the same semantics or not.

Semantic Textual Similarity Benchmark (STS-B) (Cer et al., 2017) is a dataset of paired-up media captions, news headlines, and sentences from natural language data that are given a similarity score from one to five by a human annotator. From a formal perspective, this is a regression problem where the output ranges between one and five. In a similar fashion to the last two tasks, this task can provide insights into the relevance and coherence of a response to its preceding utterance by assessing its semantic similarity.

Question Natural Language Inference (QNLI) (Wang et al., 2019) dataset is a re-adapted version of the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). The original dataset contains question-paragraph pairs, where an excerpt of the paragraph is an answer to the question. Wang et al. (Wang et al., 2019) convert it such that a question is paired up with each sentence from the context paragraph. Only the sentence with the answer for the questions has a label for textual entailment; the rest do not. The question is a hypothesis that could entail the sentence or not. It is treated as a relevance ranking problem, where a question can be more relevant to a sentence than others. Regarding dialogue quality, such a task can help with a response's relevancy assessment more straightforwardly than MRPC, QQP, and STS-B.

Recognizing Textual Entailment (RTE) datasets (Wang et al., 2019) consist of series of challenges: RTE1 (Dagan et al., 2005), RTE2 (Bar-Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009). Pairs of sentences have been sampled from news and Wikipedia articles, which have been marked, similarly to QNLI, as textual entailment or no textual entailment[3], a binary classification problem. In a similar fashion to QNLI, RTE can be used to determine the relevancy of a response to an utterance. However, unlike QNLI, RTE does so for general statements rather than just questions.

Multi-Genre Natural Language Inference Corpus (MNLI) (Williams et al., 2018) is a compilation of sentence couples collected via crowdsourcing that have been annotated for textual entailment, similarly to QNLI and RTE. However, MNLI does that as a three-class classification problem - textual entailment, contradiction, and neutrality. The task is not used for the paper due to the lack of a straightforward mapping of those three classes to an ordinal/continuous variable like a dialogue quality score.

Winograd Schema Challenge (WNLI) (Levesque et al., 2012) aims at reading comprehension where a system must gain an understanding of a sentence with a pronoun and then choose the suitable referent from a list of choices. Due to its nature, this task is not relevant and not used for this work.

4 Methodology

4.1 Dialogue Datasets

To evaluate the ability of a deep-learning model trained on GLUE to indicate the quality of dialogues, we use the English datasets (TopicalChat, PersonaChat) provided by Mehri et al (2020). They train a few different dialogue system models and use different sequence generation techniques to generate responses for certain dialogue contexts. The researchers then evaluate 660, in total, dialogue contexts and responses according to six criteria: *Understandable*, *Natural*, *Maintains Context*, *Interesting*, *Uses Knowledge*, and *Overall Quality*. For a complete description of the metrics mentioned above and further details about the dataset, we forward the reader to the original work of Mehri et al. (2020).

4.2 BERT as a Proxy Indicator for Dialogue Quality

Since the GLUE benchmark is about *general language understanding*, we are interested to know

[2] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

[3] Originally, there were two additional labels: neutral and contradiction. However, Wang et al. converted the two classes to no textual entailment (Wang et al., 2019).

whether a model trained on it can indicate the quality of the dialogue. To conduct the investigation, we use BERT (Devlin et al., 2019) and its fine-tuned models on the GLUE benchmark (Wolf et al., 2019; Morris et al., 2020). We use the version with 110M parameters. For each investigated GLUE task, there is a separate copy of the whole model trained to solve that specific problem. While we did not train the models ourselves, the inference is less demanding. It takes about 30 minutes on a laptop with an eight-generation Intel i7 CPU.

For encoding the text sequence, we use BERT, a pre-trained bidirectional transformer encoder language model. The pre-training has been done using two unsupervised tasks: masked language modeling and next sentence prediction. This way, it can learn a contextualized semantic representation of the input text usable for downstream tasks. BERT can create a vector encoding for a whole sequence by always inserting a control token, $[CLS]$, at the beginning. For the case of pair-wise sentence tasks, e.g., next sentence prediction, it uses an additional control token, $[SEP]$, between the two sentences to distinguish them.

When fine-tuned for a specific task, the pre-trained language model weights are reused. In addition, a layer is added to act as a transformation from BERT's semantic representation to the space of the target variable, e.g., the classes of RTE or CoLA.

4.3 Scoring

For obtaining model predictions, the dialogue data is provided as input in three possible ways: 1. single utterance, 2. a dialogue context and a response, or 3. related facts to a conversation and a response. Depending on the GLUE task, the model can give **four different types of output scores**:

Single-sentence classification output provides softmax output for CoLA and SST-2. Given the contextualized semantic representation of a single utterance from the dialogue U the probability whether it is linguistically acceptable or with a positive sentiment is:

$$P_r(c_{task}|U) = \text{softmax}(W_{task}^T \cdot U), \quad task \in \{\text{CoLA}, \text{SST-2}\} \tag{1}$$

where W are the task-specific weights, c is the output class for the target task.

Pairwise text similarity outputs a similarity score, for the STS-B task, between a pair of a context or fact and a target response from the same dialogue C (or F for a fact) and R, concatenated and jointly encoded by BERT as U:

$$Sim(U) = (W_{\text{STS-B}}^T \cdot U) \tag{2}$$

W are the weights specific to STS-B, and U is the concatenation of a dialogue context or fact with a target response.

Pair-wise text classification is used for the three relevant tasks of RTE, QQP, and MRPC. It functions in the same manner as single-sentence classification, with one difference. Two, instead of one, sequences are used as input to the model. The dialogue context or fact and the target response are concatenated. Between the two, a special token is inserted to signify that the input sequence has two components:

$$P_r(c_{task}|U) = \text{softmax}(W_{task}^T \cdot U), \quad task \in \{\text{RTE}, \text{QQP}, \text{MRPC}\} \tag{3}$$

Pairwise ranking finds its application in the QNLI task. Likewise to pairwise text similarity, The dialogue context or fact and the target response are concatenated C (or F) R from the same dialogue are encoded as one U to calculate a relevance score:

$$Rel(U) = g(W_{QNLI}^T \cdot U), \quad g(x) = \frac{e^x}{e^x + 1} \tag{4}$$

After model predictions are made on all utterances and sequential pairs of those across all tasks, the outputs have been rescaled between 0 and 1 for each GLUE task independently, as well as the scores given by the human annotators.

$$x'_{TASK} = \frac{x_{TASK} - min(x_{TASK})}{max(x_{TASK}) - min(x_{TASK})} \tag{5}$$

Finally, similarly as Mehri et al. (2020), we train a regression that combines all the scores in one overall score:

$$y_{overall\_score} = b + \sum_{i=0}^{GLUE} w_i \cdot x_i \tag{6}$$

5 Evaluation

Here, we analyze the dialogue datasets (Mehri and Eskénazi, 2020) for possible relations between the GLUE task predictions and the annotator scores.

5.1 Baseline: UnSupervised and Reference free (USR) evaluation metric

To bring the results into context, we compare our results to the work of Mehri et al. (2020). Their approach is reference-free and unsupervised. So, it acts as a baseline against which we compare the method proposed in this work. The algorithm has three components.

The first component, RoBERTa (Liu et al., 2019b), is fine-tuned on either PersonaChat (Zhang et al., 2018) or Topical-Chat (Gopalakrishnan et al., 2019). A concatenation of the input dialogue context and the target response is provided to its masked language modelling (MLM) objective. The tokens in the response part are iteratively replaced. In the end, the approach provides a probability score for the whole target sequence that indicates its fluency given the dialogue context. It is referred to as *USR-MLM*.

The second component again uses RoBERTa as its foundation. However, this time, it is fine-tuned on the Ubuntu Corpus (Lowe et al., 2015) to perform dialogue retrieval using negative sampling. It is trained to distinguish between the proper response of a given context and a randomly sampled one. Mehri et al. (2020) report that this metric is appropriate for evaluating Maintains Context, Interesting, and, Uses Knowledge. They refer to it as *USR-MLM* $(x = c)$ or *USR-MLM* $(x = f)$ for calculating it against the dialogue context or dialogue facts, respectively.

Finally, the third component is a combination of the other two. Mehri et al. (2020) propose using a regression model to obtain one single score based on two separate metrics. This enables measuring the overall quality of a conversation. It is referred to as only *USR*.

While Mehri et al. (2020) report turn- and system-level correlation scores. We benchmark only against turn-level scores due to a lack of detail of how the system-level ones are calculated.

5.2 Quantitative Assessment

In Tables 1 and 2, we present the correlation analysis between the automated quality metrics and human annotator scores.

In almost all of the criteria, the combined proxy indicators via linear regression outperform the combined USR metric and its best-performing components. Whereas, in the few cases where USR performs better than the proxy indicators, it is within a minor relative difference.

Looking at the *Understandable* and *Natural* criteria, we see that CoLA as a single proxy indicator can weakly infer the two measures on the TopicalChat dataset. However, it is outperformed by STSB and MRPC in PersonaChat, which suggests that the dialogues have a different nature, that involves context more strongly. This difference is also visible in the weaker performance of USR-MLM for *Understandable* and the shift to context-based USR-DR for *Natural*.

Maintains Context is the only criterion where USR outperforms the proxy indicators. Among the proxy indicators, Semantic Textual Similarity Benchmark (STSB) is the best performer, suggesting that some partial semantic overlap between context and response is necessary to model a dialogue's cohesiveness. Although, it is common sense that a reply does not need to have a high degree of semantic overlap with its context. Ultimately, the context-based USR-DR is the best-performing measure. We contribute its performance to the fact that it has been trained on dialogue data to distinguish between a correct and randomly sampled response.

We turn our attention to the *Interesting* quality measure, where USR struggles on the PersonaChat dataset. The linear regression of the proxy indicators outperforms the rest by a considerable margin. It is curious to see that the calculated STSB against the conversation data has a relatively higher correlation score. This performance suggests that responses that used the facts from the dialogue were also considered as engaging, i.e., there is an overlap between the criteria *Interesting* and *Uses Knowledge*. Aside from that, we recommend using Recognizing Textual Entailment (RTE) to indicate the interestingness of dialogue using only its context. Our results show a weak correlation with Pearson's and Spearman's coefficients ranging from 0.11 to 0.21.

The lastly mentioned metric is also the best performer for the latter criterion. Furthermore, the fact-based STSB that is compared against *Uses Knowledge* delivers the highest correlation score among all metrics. Thus, a kind of semantic similarity measure can be very indicative of whether a

knowledge base is mentioned in a conversation or not.

The linear regression of all proxy indicators appears as the most consistent performer delivering the highest scores among several specific criteria and for *Overall* one except for the context-based USR-DR, which has a higher Spearman correlation score.

All of the correlation coefficients for all pairs of predictors and human annotator criteria are available in Appendix A.

TopicalChat		
Metric	**Pearson**	**Spearman**
Understandable		
USR-MLM	0.3268	0.3264
USR	0.3152	0.2932
CoLA	0.2458	0.2341
Lin-Reg (all)	**0.3420**	**0.3390**
Natural		
USR-MLM	0.3254	0.3370
USR	0.3037	0.2763
CoLA	0.2069	0.1677
Lin-Reg (all)	**0.3357**	**0.3130**
Maintains Context		
USR-DR (x=c)	0.3650	0.3391
USR	**0.3769**	**0.4160**
STSB	0.2350	0.2340
Lin-Reg (all)	0.3489	0.3409
Interesting		
USR-DR (x=c)	0.4877	0.3533
USR	0.4645	0.4555
STSB (fact)	0.4147	0.4103
Lin-Reg (all)	**0.5335**	**0.5364**
Uses Knowledge		
USR-DR (x=f)	0.4468	0.2220
USR	0.3353	0.3175
STSB (fact)	0.4808	0.4522
Lin-Reg (all)	**0.5119**	**0.5295**
Overall		
USR-DR (x=c)	0.3245	0.4068
USR	0.4192	0.4220
STSB (fact)	0.3324	0.3220
Lin-Reg (all)	**0.4974**	**0.4877**

Table 1: Turn-level correlation results based on the sample dialogues from the TopicalChat dataset. The USR metrics are from the original work of Mehri et al (2020). Only the best performing metrics are shown in the table. All of the correlation coefficients are with a statistical significance of $p < 0.05$.

PersonaChat		
Metric	**Pearson**	**Spearman**
Understandable		
USR-MLM	0.1186	0.1313
USR	**0.1324**	**0.1241**
STSB	0.1286	0.1159
Lin-Reg (all)	0.1214	0.1218
Natural		
USR-DR (x=c)	0.2291	0.1733
USR	**0.2430**	0.1862
MRPC	0.1794	**0.2410**
Lin-Reg (all)	0.1728	0.2044
Maintains Context		
USR-DR (x=c)	**0.5625**	0.6021
USR	0.5280	**0.6065**
STSB	0.3620	0.3463
Lin-Reg (all)	0.4029	0.3707
Interesting		
USR-DR (x=c)	0.2634	0.0606
USR	0.0171	0.0315
STSB (fact)	**0.3419**	**0.3378**
Lin-Reg (all)	0.3272	0.3306
Uses Knowledge		
USR-DR (x=c)	0.6309	0.4508
USR	0.3177	0.4027
STSB (fact)	**0.7329**	**0.7173**
Lin-Reg (all)	0.5921	0.5898
Overall		
USR-DR (x=c)	0.4814	**0.6087**
USR	0.4693	0.4115
STSB (fact)	0.3742	0.3898
Lin-Reg (all)	**0.5290**	0.5382

Table 2: Turn-level correlation results based on the sample dialogues from the PersonaChat dataset. The USR metrics are from the original work of Mehri et al (2020). Only the best performing metrics are shown in the table. All of the correlation coefficients are with a statistical significance of $p < 0.05$.

5.3 Ablation Study

We investigate four configurations for using a different subset of the proxy indicators to calculate a combined score using linear regression and check the correlation coefficients against the various dialogue criteria:

- **Lin-Reg (single)** - a linear regression combining only the single-sentence GLUE tasks applied on the target response - CoLA, SST-2
- **Lin-Reg (context)** - a linear regression combining only the pair-wise sentence GLUE tasks that model the dialogue context, and the

TopicalChat		
Metric	**Pearson**	**Spearman**
Understandable		
Lin-Reg (single)	0.2542	0.2470
Lin-Reg (context)	0.1664	0.1638
Lin-Reg (fact)	0.2572	0.2362
Lin-Reg (all)	0.3420	0.3390
Natural		
Lin-Reg (single)	0.2148	0.1853
Lin-Reg (context)	0.1986	0.1972
Lin-Reg (fact)	0.2244	0.1805
Lin-Reg (all)	0.3357	0.3130
Maintains Context		
Lin-Reg (single)	*0.0469*	*0.0197*
Lin-Reg (context)	0.2859	0.2946
Lin-Reg (fact)	0.2272	0.1921
Lin-Reg (all)	0.3489	0.3409
Interesting		
Lin-Reg (single)	0.1483	*0.0881*
Lin-Reg (context)	0.3884	0.4008
Lin-Reg (fact)	0.4358	0.4078
Lin-Reg (all)	0.5335	0.5364
Uses Knowledge		
Lin-Reg (single)	0.0699	0.0377
Lin-Reg (context)	0.2455	0.2751
Lin-Reg (fact)	0.5517	0.5182
Lin-Reg (all)	0.5119	0.5295
Overall		
Lin-Reg (single)	0.1432	0.1138
Lin-Reg (context)	0.3492	0.3587
Lin-Reg (fact)	0.3897	0.3482
Lin-Reg (all)	0.4974	0.4877

Table 3: Turn-level correlation results for different mixtures of proxy indicators based on the sample dialogues from the TopicalChat dataset. All of the correlation coefficients except the ones with *italics* have a statistical significance of $p < 0.05$.

PersonaChat		
Metric	**Pearson**	**Spearman**
Understandable		
Lin-Reg (single)	*0.0643*	*0.0603*
Lin-Reg (context)	0.1626	0.1345
Lin-Reg (fact)	*0.0255*	*0.0328*
Lin-Reg (all)	0.1214	0.1218
Natural		
Lin-Reg (single)	*-0.0285*	*0.0302*
Lin-Reg (context)	0.2033	0.2160
Lin-Reg (fact)	*0.0546*	*0.0319*
Lin-Reg (all)	0.1728	0.2044
Maintains Context		
Lin-Reg (single)	*0.0974*	*0.1012*
Lin-Reg (context)	0.4178	0.3981
Lin-Reg (fact)	0.1783	0.1110
Lin-Reg (all)	0.4029	0.3707
Interesting		
Lin-Reg (single)	0.1675	0.1597
Lin-Reg (context)	0.2185	0.2216
Lin-Reg (fact)	0.3446	0.3412
Lin-Reg (all)	0.3272	0.3306
Uses Knowledge		
Lin-Reg (single)	*0.0464*	*0.0644*
Lin-Reg (context)	0.1909	0.1916
Lin-Reg (fact)	0.6959	0.7020
Lin-Reg (all)	0.5921	0.5898
Overall		
Lin-Reg (single)	0.1216	0.1263
Lin-Reg (context)	0.3975	0.3802
Lin-Reg (fact)	0.3990	0.4135
Lin-Reg (all)	0.5290	0.5382

Table 4: Turn-level correlation results for different mixtures of proxy indicators based on the sample dialogues from the PersonaChat dataset. All of the correlation coefficients except the ones with *italics* have a statistical significance of $p < 0.05$.

target response - MRPC, QQP, STSB, QNLI, RTE

- **Lin-Reg (fact)** - a linear regression combining only the pair-wise sentence GLUE tasks that model the dialogue facts, and the target response - MRPC, QQP, STSB, QNLI, RTE

- **Lin-Reg (all)** - a linear regression combining all of GLUE tasks that model the dialogue context, fact, and the target response

The combination of single sentence tasks shows signs of capability only on the criteria which can be evaluated utterance-wise, *Understandable*, *Natural*, and *Interesting*. While in the others, there is a drop in correlation coefficients and statistical significance, which agrees with general intuition. The single-sentence tasks cannot model dialogue quality metrics that require a view beyond the single utterances.

Turning to *Maintains Context*, we see the inverse perspective. The pair-wise sentence proxy indicators applied to the dialogue context, and target response demonstrate the best ability, while the single sentence is the worst. Furthermore, the observation is partially supported by the pair-wise tasks applied to the dialogue facts.

In regards to *Interesting*, it is evident that the pair-wise tasks outperform the single-sentence ones since context dictates what is engaging in a conversation rather than the single utterances.

Moreover, the fact-based pair-wise proxy indicators demonstrate their strong ability to model the *Uses Knowledge* criterion since these are the only automatic metrics that have access to the fact information. In comparison, the others underperform since they are not evaluated against the relevant data.

Finally, it is evident that to calculate an *Overall* score, one needs to use all of the proxy indicators. All of the subset combinations perform worse than the linear regression combining all of the metrics. Moreover, we see how the correlation improves for the combined score regarding the specific criteria like *Maintains Context*, and *Interesting*.

5.4 GLUE Predictor Feature Importance

In Figure 1, we present the inferred weights of the single GLUE predictors via linear regression.

It is immediately evident that in both datasets, the single sentence tasks, CoLA and SST, have

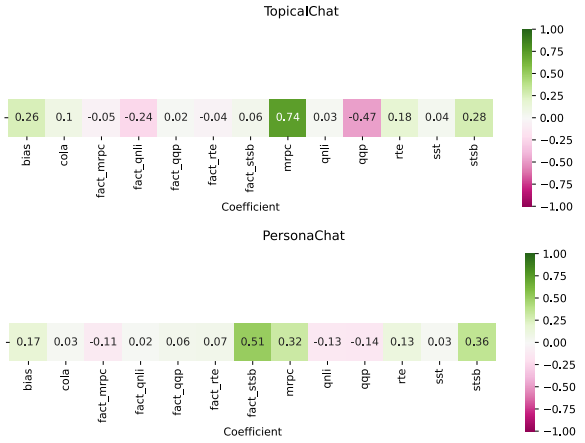

Figure 1: The weights as inferred by the linear regression **Lin-Reg (all)**) for each of the single GLUE predictors.

an insignificant influence on the prediction of the overall quality score.

Semantic overlap between the utterances via STS-B and MRPC plays in both cases a significant role. However, in TopicalChat, the latter of the two has an even more substantial part. The trivia-like nature of the conversations explains the behavior. The significant scores of QQP and QNLI between facts and conversation utterances support the observation.

Looking at the influence of knowledge-base-related predictors, we see that in PersonaChat, it is essential to have semantic similarity (STSB) with the knowledge base facts, i.e., that the dialogue systems use the personal traits in the conversation.

5.5 Error Analysis

In Appendix B, we present regression plots with 95% confidence intervals in order to inspect for errors. We present the following conclusions:

- The linear regression on all scores has a decent general performance. Its weakness is the lower-end spectrum of the human-annotator overall quality criteria. There is a higher score variance, i.e., higher disagreement between the annotators.

- STS-B performs well on the "clear-cut" samples where knowledge is used or not. However, on borderline cases, where annotators disagree, i.e., some say knowledge is used and others not, it performs worse.

- CoLA performs excellently on the samples that were marked as Understandable by all an-

notators. As the scores for understandability decrease, so does the inter-annotator agreement. Hence, also the performance of CoLA.

Overall, it appears that the approach suffers the most when there is a high disagreement between the annotators, which are on the lower end of the human annotator scoring.

The USR dataset includes information about the annotators in the form of nicknames. Based on those, one can assume that they were non-native English speakers with various backgrounds. Hence, there is a low inter-annotator agreement on "Understandable" and "Natural." For example, native speakers of a Romance and a Slavic language are more likely to disagree on these two criteria. Furthermore, it is also confirmed by the higher variance in the annotator score on the lower spectrum of CoLA predictions, i.e., annotators agree well, what understandable language is, but not the opposite.

6 Conclusion

This work considered a model trained on GLUE as a proxy indicator for the quality of knowledge-grounded dialogues offering different perspectives on dialogue quality criteria. It does not need any references or supervision and can outperform other competing approaches like USR (Mehri and Eskénazi, 2020). Pearson's and Spearman's correlation coefficients suggest that single proxy indicators and their various combinations via linear regression can infer dialogue quality either on specific criteria or in general. This composable nature can be used to tune the approach to focus more on particular criteria than others.

While one might be concerned that using the approach might offer an advantage to dialogue systems incorporating BERT, we think it poses little to no risk. BERT is an encoder approach and is considered uncommon for sequence generation applications. Hence, the risk of bias is reasonably low. In addition, one could also use any other base model architecture for training GLUE predictors.

The model has no training or fine-tuning that is specifically geared towards dialogues. However, we showed that lack of exposure to conversational data could be problematic for metrics like *Maintains Context*. Hence, we set as future work to investigate additional pre-training on dialogue data similarly as Mehri et al. (2020), but also considering other proxy indicators like DialogueNLI (Welleck et al., 2019), which frame the natural language inference task in a conversational setting.

Finally, while we used separately trained instances of BERT for each of the GLUE tasks, one could also consider using a multi-tasking method. For example, Liu et al (2019a) present Multi-Task Deep Neural Networks (MT-DNN) that employ a single instance of BERT for all GLUE tasks. We believe using multi-tasking and BERT together would make its application in a productive environment much more effortless, since model weights are to a greater extent shared between the tasks.

7 Acknowledgements

We acknowledge the support of the following projects: SPEAKER (FKZ 01MK20011A), JOSEPH (Fraunhofer Zukunftsstiftung), H2020 Cleopatra (GA 812997), ML2R (FKZ 01 15 18038 A/B/C), ScaDS.AI (BMBF 01IS18026A) and TAILOR (GA 952215).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

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

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. *arXiv preprint arXiv:1906.09308*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1891–1895. ISCA.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4487–4496. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Varvara Logacheva, Mikhail Burtsev, Valentin Malykh, Vadim Polulyakh, and Aleksandr Seliverstov. 2018. Convai dataset of topic-oriented human-to-chatbot dialogues. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 47–57. Springer.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294. The Association for Computer Linguistics.

Shikib Mehri and Maxine Eskénazi. 2020. USR: an unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 681–707. Association for Computational Linguistics.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.

Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck. 2020a. Language model transformers as evaluators for open-domain dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6797–6808. International Committee on Computational Linguistics.

Rostislav Nedelchev, Ricardo Usbeck, and Jens Lehmann. 2020b. Treating dialogue quality evaluation as an anomaly detection problem. In *Proceedings of The 12th Language Resources and Evalua-*

tion Conference, pages 501–505, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Trans. Assoc. Comput. Linguistics*, 8:810–827.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3731–3741. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

A Complete correlation scores for all predictors

We present complete tables with correlation scores of all pairs of predictors and human annotator scores. In Tables 5, and 6 are the correlation scores for the single GLUE tasks. Furthermore, Tables 7, and 8 present the correlation coefficients for on the GLUE predictions of the knowledge base facts and the dialogue utterances. Finally, Tables 9, and 10 show the correlation scores for the various combinations of the GLUE predictors using linear regression.

TopicalChat				
Predictor-Criteria	**Pearson's r**	$p <$	**Spearman's ρ**	$p <$
cola-Understandable	0.2458	0.0001	0.2341	0.0001
cola-Natural	0.2069	0.0001	0.1677	0.0014
cola-Maintains Context	0.0449	0.3959	0.0119	0.8226
cola-Engaging	0.1518	0.0039	0.0935	0.0765
cola-Uses Knowledge	0.0727	0.1686	0.0481	0.3623
cola-Overall	0.1418	0.0070	0.1136	0.0312
sst-Understandable	0.1253	0.0173	0.1114	0.0346
sst-Natural	0.1107	0.0358	0.0826	0.1176
sst-Maintains Context	0.0260	0.6225	-0.0064	0.9041
sst-Engaging	0.0146	0.7825	-0.0328	0.5346
sst-Uses Knowledge	-0.0006	0.9906	-0.0517	0.3280
sst-Overall	0.0471	0.3731	0.0139	0.7924
mrpc-Understandable	0.1216	0.0210	0.0890	0.0918
mrpc-Natural	0.1366	0.0095	0.1171	0.0264
mrpc-Maintains Context	0.2083	0.0001	0.2131	0.0001
mrpc-Engaging	0.0985	0.0619	0.0823	0.1191
mrpc-Uses Knowledge	-0.0395	0.4545	-0.0266	0.6147
mrpc-Overall	0.1419	0.0070	0.1258	0.0170
qnli-Understandable	-0.0076	0.8864	0.0062	0.9069
qnli-Natural	-0.0095	0.8571	-0.0032	0.9515
qnli-Maintains Context	-0.0078	0.8824	-0.0015	0.9768
qnli-Engaging	0.1409	0.0074	0.1538	0.0034
qnli-Uses Knowledge	0.1382	0.0086	0.1509	0.0041
qnli-Overall	0.0853	0.1060	0.0952	0.0711
qqp-Understandable	-0.0311	0.5569	-0.0369	0.4858
qqp-Natural	-0.0510	0.3346	-0.0142	0.7879
qqp-Maintains Context	-0.0173	0.7439	0.0529	0.3173
qqp-Engaging	-0.0845	0.1095	-0.0910	0.0848
qqp-Uses Knowledge	-0.1103	0.0365	-0.1352	0.0102
qqp-Overall	-0.0751	0.1548	-0.0708	0.1804
rte-Understandable	0.0598	0.2577	0.0758	0.1510
rte-Natural	0.0833	0.1147	0.0936	0.0761
rte-Maintains Context	-0.0131	0.8043	-0.0419	0.4282
rte-Engaging	0.2024	0.0001	0.2116	0.0001
rte-Uses Knowledge	0.2478	0.0001	0.2523	0.0001
rte-Overall	0.1619	0.0021	0.1554	0.0031
stsb-Understandable	0.0343	0.5160	0.0473	0.3711
stsb-Natural	0.0270	0.6094	0.0430	0.4158
stsb-Maintains Context	0.2350	0.0001	0.2340	0.0001
stsb-Engaging	0.1457	0.0056	0.1704	0.0012
stsb-Uses Knowledge	-0.0056	0.9150	0.0360	0.4962
stsb-Overall	0.1129	0.0322	0.1429	0.0066

Table 5: Correlation scores between the GLUE tasks on the conversation utterances and the human annotator scores and their respective p-values on the **TopicalChat** dataset.

PersonaChat				
Predictor-Criteria	**Pearson's r**	$p <$	**Spearman's ρ**	$p <$
cola-Understandable	0.0318	0.5828	0.0673	0.2451
cola-Natural	0.0838	0.1475	-0.0309	0.5945
cola-Maintains Context	-0.0862	0.1365	-0.1935	0.0008
cola-Engaging	-0.0665	0.2510	-0.1568	0.0065
cola-Uses Knowledge	-0.0190	0.7425	-0.1403	0.0150
cola-Overall	-0.0252	0.6635	-0.1931	0.0008
sst-Understandable	0.0743	0.1996	0.0723	0.2119
sst-Natural	-0.0064	0.9119	0.0294	0.6123
sst-Maintains Context	0.0760	0.1890	0.0988	0.0875
sst-Engaging	0.1530	0.0080	0.1242	0.0315
sst-Uses Knowledge	0.0422	0.4663	-0.0034	0.9531
sst-Overall	0.1172	0.0424	0.1068	0.0647
mrpc-Understandable	0.0857	0.1385	0.1098	0.0574
mrpc-Natural	0.1794	0.0018	0.2410	0.0001
mrpc-Maintains Context	0.3129	0.0001	0.3684	0.0001
mrpc-Engaging	-0.1266	0.0284	0.0695	0.2301
mrpc-Uses Knowledge	-0.0656	0.2574	-0.0112	0.8468
mrpc-Overall	0.1959	0.0006	0.2576	0.0001
qnli-Understandable	-0.1356	0.0188	-0.1434	0.0129
qnli-Natural	-0.1821	0.0015	-0.2058	0.0003
qnli-Maintains Context	-0.3795	0.0001	-0.3982	0.0001
qnli-Engaging	0.0163	0.7780	0.0318	0.5832
qnli-Uses Knowledge	-0.0430	0.4580	-0.0490	0.3981
qnli-Overall	-0.2553	0.0001	-0.2434	0.0001
qqp-Understandable	0.0529	0.3613	0.0830	0.1514
qqp-Natural	0.1071	0.0639	0.1857	0.0012
qqp-Maintains Context	0.1646	0.0043	0.3472	0.0001
qqp-Engaging	-0.3205	0.0001	-0.0071	0.9029
qqp-Uses Knowledge	-0.1725	0.0027	0.0208	0.7198
qqp-Overall	0.0276	0.6345	0.2125	0.0002
rte-Understandable	-0.0519	0.3704	-0.0976	0.0916
rte-Natural	-0.0710	0.2200	-0.1184	0.0404
rte-Maintains Context	-0.2789	0.0001	-0.2999	0.0001
rte-Engaging	0.1131	0.0503	0.1269	0.0280
rte-Uses Knowledge	0.0752	0.1939	0.0827	0.1531
rte-Overall	-0.0842	0.1459	-0.0766	0.1860
stsb-Understandable	0.1286	0.0259	0.1159	0.0448
stsb-Natural	0.1140	0.0486	0.1317	0.0225
stsb-Maintains Context	0.3620	0.0001	0.3463	0.0001
stsb-Engaging	0.0889	0.1242	0.0805	0.1645
stsb-Uses Knowledge	0.0988	0.0877	0.0828	0.1525
stsb-Overall	0.2591	0.0001	0.2396	0.0001

Table 6: Correlation scores between the GLUE tasks on the conversation utterances and the human annotator scores and their respective p-values on the **PersonaChat** dataset.

TopicalChat				
Predictor-Criteria	**Pearson's r**	$p <$	**Spearman's ρ**	$p <$
fact_mrpc-Understandable	0.1357	0.0100	0.1935	0.0002
fact_mrpc-Natural	0.0564	0.2859	0.1186	0.0244
fact_mrpc-Maintains Context	0.0827	0.1174	0.1981	0.0002
fact_mrpc-Engaging	0.2017	0.0001	0.3052	0.0001
fact_mrpc-Uses Knowledge	0.3162	0.0001	0.3839	0.0001
fact_mrpc-Overall	0.1749	0.0009	0.2647	0.0001
fact_qnli-Understandable	-0.2597	0.0001	-0.2355	0.0001
fact_qnli-Natural	-0.2419	0.0001	-0.1981	0.0002
fact_qnli-Maintains Context	-0.2239	0.0001	-0.1842	0.0004
fact_qnli-Engaging	-0.4034	0.0001	-0.3727	0.0001
fact_qnli-Uses Knowledge	-0.5291	0.0001	-0.5457	0.0001
fact_qnli-Overall	-0.3784	0.0001	-0.3390	0.0001
fact_qqp-Understandable	0.1656	0.0016	0.2147	0.0001
fact_qqp-Natural	0.1217	0.0209	0.1788	0.0007
fact_qqp-Maintains Context	0.1607	0.0022	0.1917	0.0003
fact_qqp-Engaging	0.3197	0.0001	0.3824	0.0001
fact_qqp-Uses Knowledge	0.4373	0.0001	0.5350	0.0001
fact_qqp-Overall	0.2683	0.0001	0.3347	0.0001
fact_rte-Understandable	-0.1823	0.0005	-0.1896	0.0003
fact_rte-Natural	-0.1512	0.0040	-0.1408	0.0075
fact_rte-Maintains Context	-0.1297	0.0138	-0.1398	0.0079
fact_rte-Engaging	-0.2565	0.0001	-0.2620	0.0001
fact_rte-Uses Knowledge	-0.3900	0.0001	-0.5263	0.0001
fact_rte-Overall	-0.2312	0.0001	-0.2360	0.0001
fact_stsb-Understandable	0.1994	0.0001	0.1999	0.0001
fact_stsb-Natural	0.1346	0.0106	0.1249	0.0178
fact_stsb-Maintains Context	0.1832	0.0005	0.1739	0.0009
fact_stsb-Engaging	0.4147	0.0001	0.4103	0.0001
fact_stsb-Uses Knowledge	0.4808	0.0001	0.4522	0.0001
fact_stsb-Overall	0.3324	0.0001	0.3220	0.0001

Table 7: Correlation scores between the GLUE tasks on the conversation utterances evaluated against the knowledge base facts and the human annotator scores and their respective p-values on the **TopicalChat** dataset.

PersonaChat				
Predictor-Criteria	**Pearson's r**	$p <$	**Spearman's ρ**	$p <$
fact_mrpc-Understandable	0.1219	0.0349	0.1938	0.0007
fact_mrpc-Natural	0.0417	0.4721	0.0550	0.3425
fact_mrpc-Maintains Context	0.0252	0.6642	-0.0532	0.3589
fact_mrpc-Engaging	0.1302	0.0241	0.0461	0.4265
fact_mrpc-Uses Knowledge	-0.0046	0.9367	-0.0571	0.3247
fact_mrpc-Overall	0.0149	0.7972	-0.0726	0.2101
fact_qnli-Understandable	-0.1256	0.0296	-0.1494	0.0095
fact_qnli-Natural	-0.0642	0.2674	-0.0584	0.3131
fact_qnli-Maintains Context	-0.1478	0.0103	-0.1014	0.0794
fact_qnli-Engaging	-0.1157	0.0453	-0.0817	0.1583
fact_qnli-Uses Knowledge	-0.2733	0.0001	-0.2613	0.0001
fact_qnli-Overall	-0.1899	0.0009	-0.1734	0.0026
fact_qqp-Understandable	0.0476	0.4113	-0.0767	0.1850
fact_qqp-Natural	0.0762	0.1881	-0.0591	0.3072
fact_qqp-Maintains Context	0.0397	0.4936	0.0774	0.1813
fact_qqp-Engaging	0.1400	0.0152	0.1365	0.0180
fact_qqp-Uses Knowledge	0.2099	0.0003	0.4352	0.0001
fact_qqp-Overall	0.1613	0.0051	0.2230	0.0001
fact_rte-Understandable	-0.0296	0.6098	-0.0914	0.1142
fact_rte-Natural	0.0289	0.6181	0.0305	0.5993
fact_rte-Maintains Context	-0.0371	0.5225	-0.0041	0.9440
fact_rte-Engaging	-0.1091	0.0591	-0.0726	0.2100
fact_rte-Uses Knowledge	-0.4052	0.0001	-0.3481	0.0001
fact_rte-Overall	-0.1232	0.0330	-0.1122	0.0522
fact_stsb-Understandable	0.0250	0.6660	0.0307	0.5961
fact_stsb-Natural	0.0302	0.6025	-0.0032	0.9555
fact_stsb-Maintains Context	0.1537	0.0077	0.0876	0.1300
fact_stsb-Engaging	0.3419	0.0001	0.3378	0.0001
fact_stsb-Uses Knowledge	0.7329	0.0001	0.7173	0.0001
fact_stsb-Overall	0.3742	0.0001	0.3898	0.0001

Table 8: Correlation scores between the GLUE tasks on the conversation utterances evaluated against the knowledge base facts and the human annotator scores and their respective p-values on the **PersonaChat** dataset.

TopicalChat				
Predictor-Criteria	**Pearson's** r	$p <$	**Spearman's** ρ	$p <$
lin-reg_pair-Understandable	0.1664	0.0015	0.1638	0.0018
lin-reg_pair-Natural	0.1986	0.0001	0.1972	0.0002
lin-reg_pair-Maintains Context	0.2859	0.0001	0.2946	0.0001
lin-reg_pair-Engaging	0.3884	0.0001	0.4008	0.0001
lin-reg_pair-Uses Knowledge	0.2455	0.0001	0.2751	0.0001
lin-reg_pair-Overall	0.3492	0.0001	0.3587	0.0001
lin-reg_fact-Understandable	0.2572	0.0001	0.2362	0.0001
lin-reg_fact-Natural	0.2244	0.0001	0.1805	0.0006
lin-reg_fact-Maintains Context	0.2272	0.0001	0.1921	0.0002
lin-reg_fact-Engaging	0.4358	0.0001	0.4078	0.0001
lin-reg_fact-Uses Knowledge	0.5517	0.0001	0.5182	0.0001
lin-reg_fact-Overall	0.3897	0.0001	0.3482	0.0001
lin-reg_single-Understandable	0.2542	0.0001	0.2470	0.0001
lin-reg_single-Natural	0.2148	0.0001	0.1853	0.0004
lin-reg_single-Maintains Context	0.0469	0.3753	0.0197	0.7094
lin-reg_single-Engaging	0.1483	0.0048	0.0881	0.0952
lin-reg_single-Uses Knowledge	0.0699	0.1855	0.0377	0.4754
lin-reg_single-Overall	0.1432	0.0065	0.1138	0.0308
lin-reg_all-Understandable	0.3420	0.0001	0.3390	0.0001
lin-reg_all-Natural	0.3357	0.0001	0.3130	0.0001
lin-reg_all-Maintains Context	0.3489	0.0001	0.3409	0.0001
lin-reg_all-Engaging	0.5335	0.0001	0.5364	0.0001
lin-reg_all-Uses Knowledge	0.5119	0.0001	0.5295	0.0001
lin-reg_all-Overall	0.4974	0.0001	0.4877	0.0001

Table 9: Correlation scores between the combined GLUE scores with linear regression and the human annotator scores and their respective p-values on the **TopicalChat** dataset.

PersonaChat				
Predictor-Criteria	**Pearson's** r	$p <$	**Spearman's** ρ	$p <$
lin-reg_pair-Understandable	0.1626	0.0047	0.1345	0.0198
lin-reg_pair-Natural	0.2033	0.0004	0.2160	0.0002
lin-reg_pair-Maintains Context	0.4178	0.0001	0.3981	0.0001
lin-reg_pair-Engaging	0.2185	0.0001	0.2216	0.0001
lin-reg_pair-Uses Knowledge	0.1909	0.0009	0.1916	0.0009
lin-reg_pair-Overall	0.3975	0.0001	0.3802	0.0001
lin-reg_fact-Understandable	0.0255	0.6606	0.0328	0.5720
lin-reg_fact-Natural	0.0546	0.3456	0.0319	0.5823
lin-reg_fact-Maintains Context	0.1783	0.0019	0.1110	0.0548
lin-reg_fact-Engaging	0.3446	0.0001	0.3412	0.0001
lin-reg_fact-Uses Knowledge	0.6959	0.0001	0.7020	0.0001
lin-reg_fact-Overall	0.3990	0.0001	0.4135	0.0001
lin-reg_single-Understandable	0.0643	0.2668	0.0603	0.2978
lin-reg_single-Natural	-0.0285	0.6226	0.0302	0.6024
lin-reg_single-Maintains Context	0.0974	0.0922	0.1012	0.0801
lin-reg_single-Engaging	0.1675	0.0036	0.1597	0.0056
lin-reg_single-Uses Knowledge	0.0464	0.4230	0.0644	0.2661
lin-reg_single-Overall	0.1216	0.0353	0.1263	0.0287
lin-reg_all-Understandable	0.1214	0.0355	0.1218	0.0350
lin-reg_all-Natural	0.1728	0.0027	0.2044	0.0004
lin-reg_all-Maintains Context	0.4029	0.0001	0.3707	0.0001
lin-reg_all-Engaging	0.3272	0.0001	0.3306	0.0001
lin-reg_all-Uses Knowledge	0.5921	0.0001	0.5898	0.0001
lin-reg_all-Overall	0.5290	0.0001	0.5382	0.0001

Table 10: Correlation scores between the combined GLUE scores with linear regression and the human annotator scores and their respective p-values on the **PersonaChat** dataset.

B Regression plots between predictions and human annotator scores

We provide regression plots with 95% confidence intervals between predictions and human annotator scores. Figures 2, and 3 show the correlation between the single GLUE predictions and the human annotator scores for TopicalChat and PersonaChat, respectively. While, Figures 4, and 5 show the correlation between the various combinations using linear regression and the human annotator scores for TopicalChat and PersonaChat, respectively. The vertical lines represent the prediction distribution for the given averaged annotator score within a 95% confidence interval. The dot signifies the mean value. For example, looking at Figure 5, subplot "lin-reg_fact | Uses Knowledge," the line overlaps well with the lowest (0) and the highest score (1), meaning that the prediction can distinguish well between when a dialogue uses knowledge or not. However, in the cases where the annotators could not agree, the predictor tends to overestimate them using knowledge since the intervals are below the regression line.

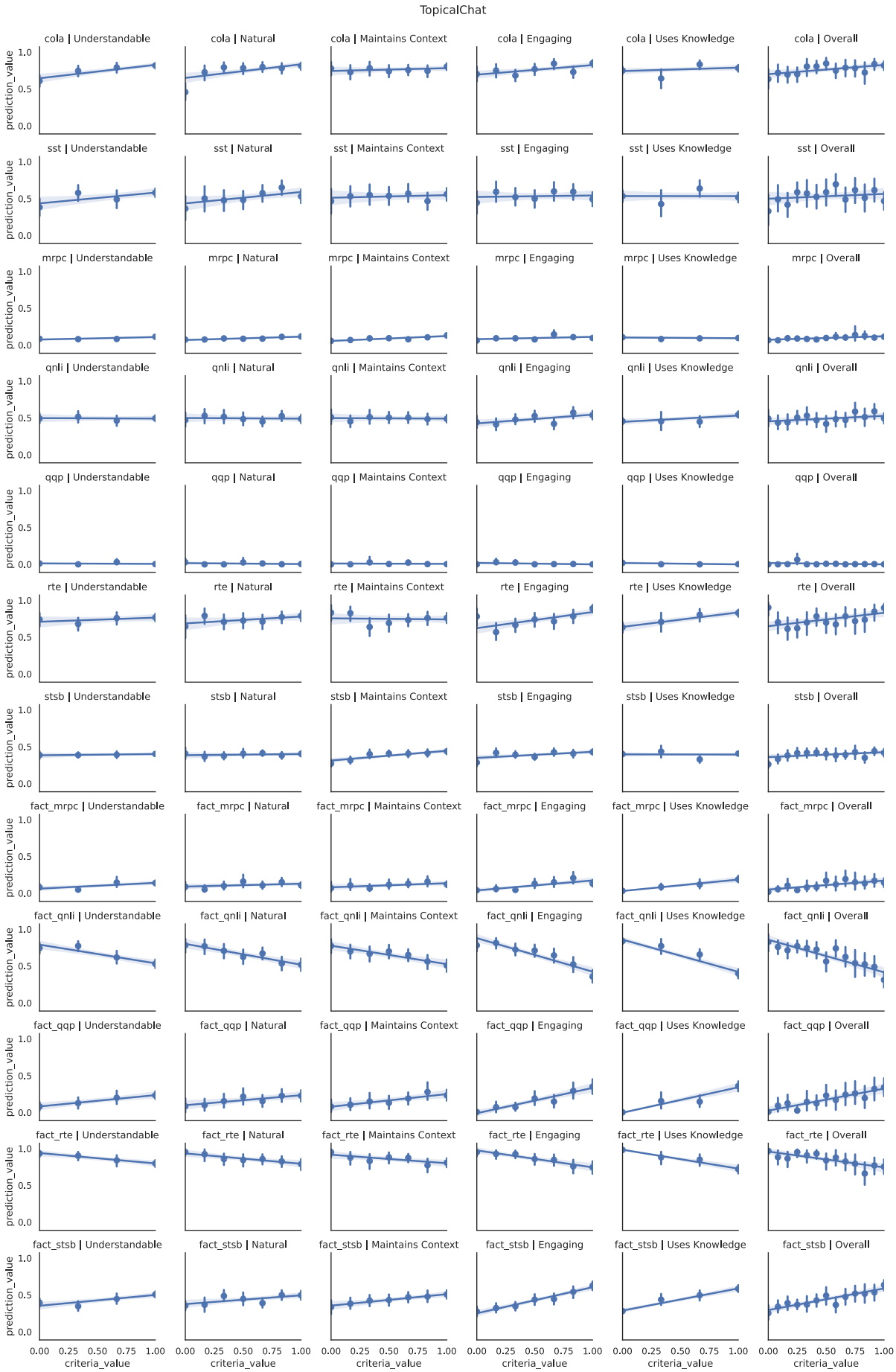

Figure 2: Regression plots between the single GLUE predictors and the human annotator scores on **TopicalChat**

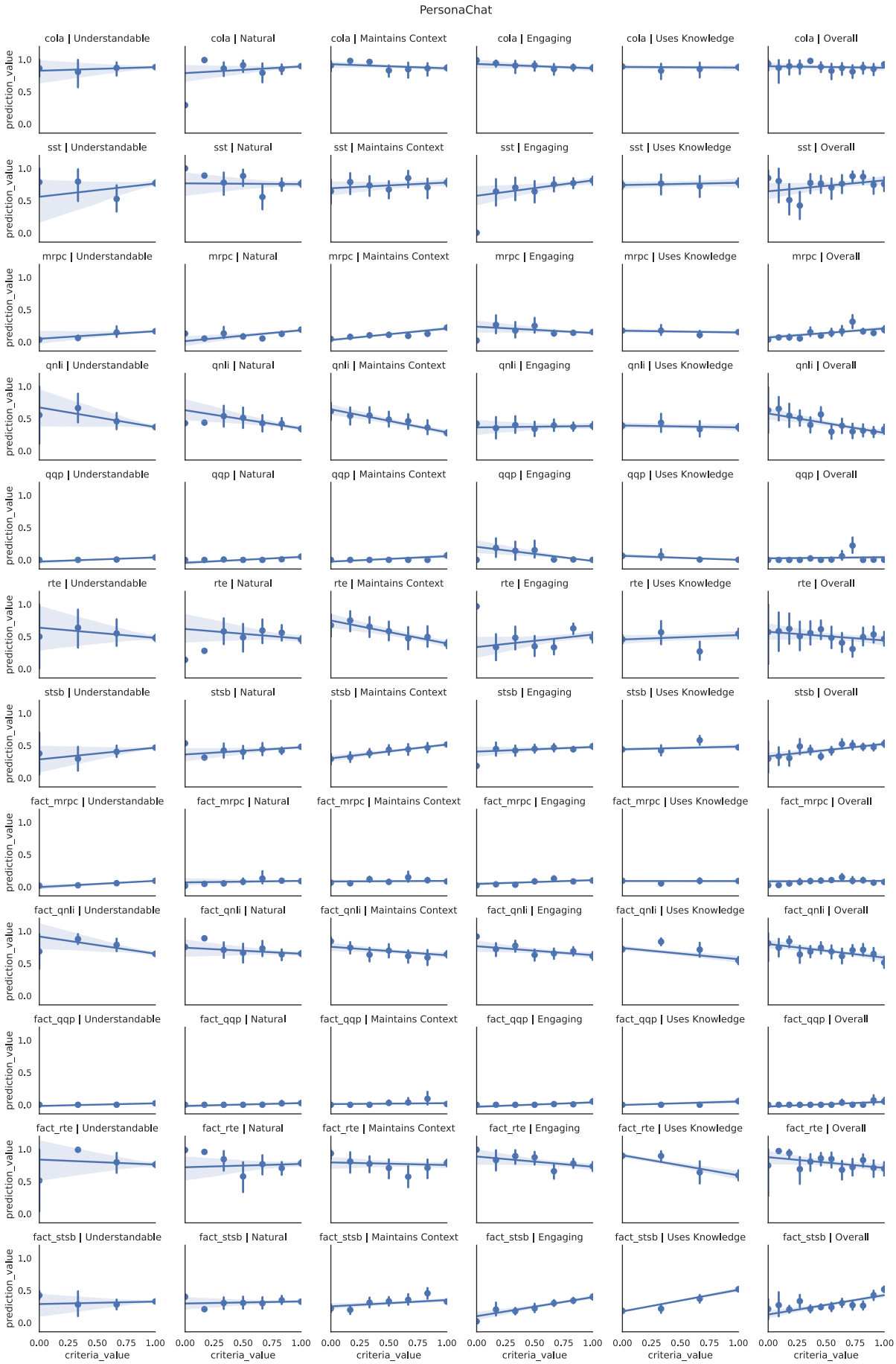

Figure 3: Regression plots between the single GLUE predictors and the human annotator scores on **PersonaChat**

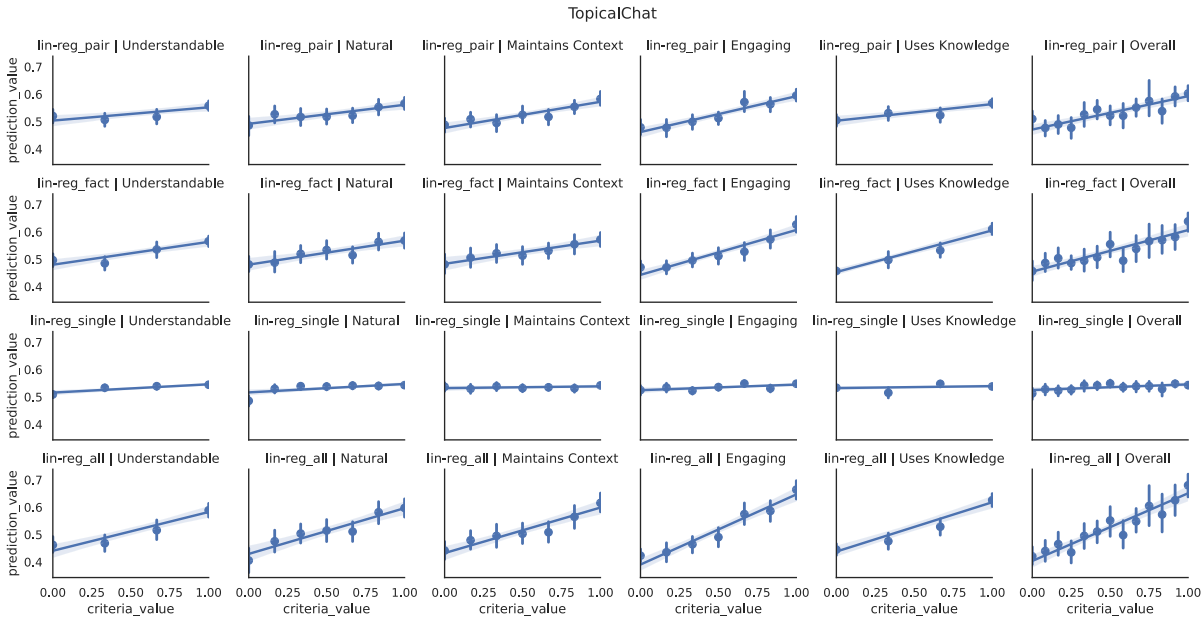

Figure 4: Regression plots between the combined linear regression predictors and the human annotator scores on **TopicalChat**

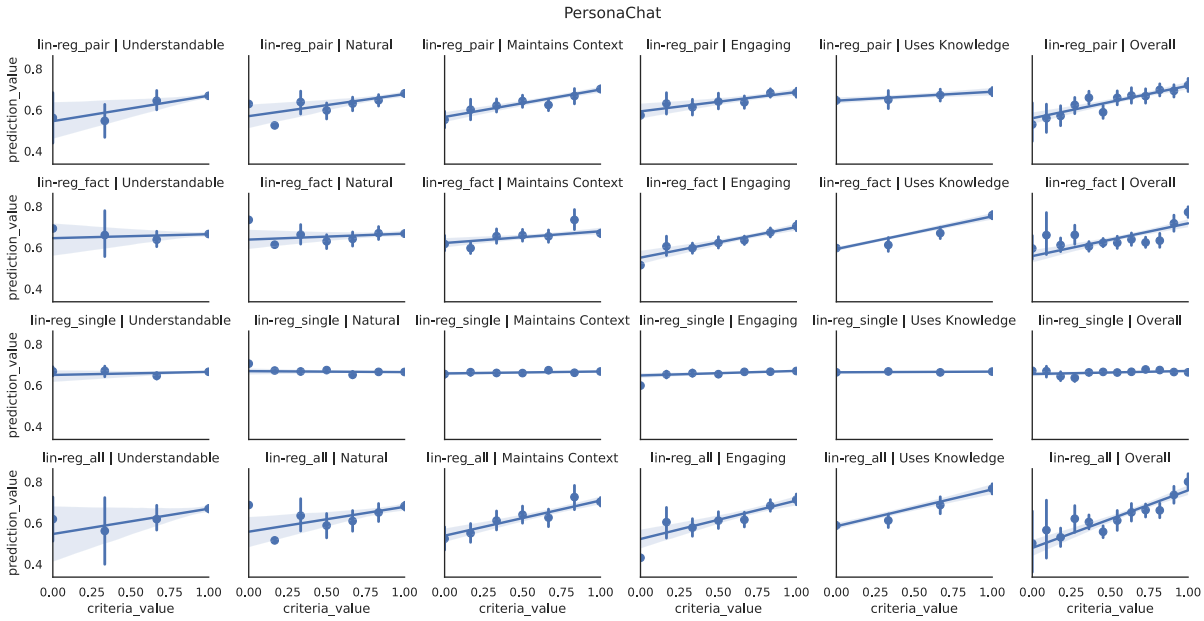

Figure 5: Regression plots between the combined linear regression predictors and the human annotator scores on **PersonaChat**