Exploring the Role of Context in Utterance-level Emotion, Act and Intent Classification in Conversations: An Empirical Study

Deepanway Ghosal[†], **Navonil Majumder**[†], **Rada Mihalcea**[△], **Soujanya Poria**[†]

[†] Singapore University of Technology and Design, Singapore
[△] University of Michigan, USA
deepanway_ghosal@mymail.sutd.edu.sg
{navonil_majumder, sporia}@sutd.edu.sg
mihalcea@umich.edu

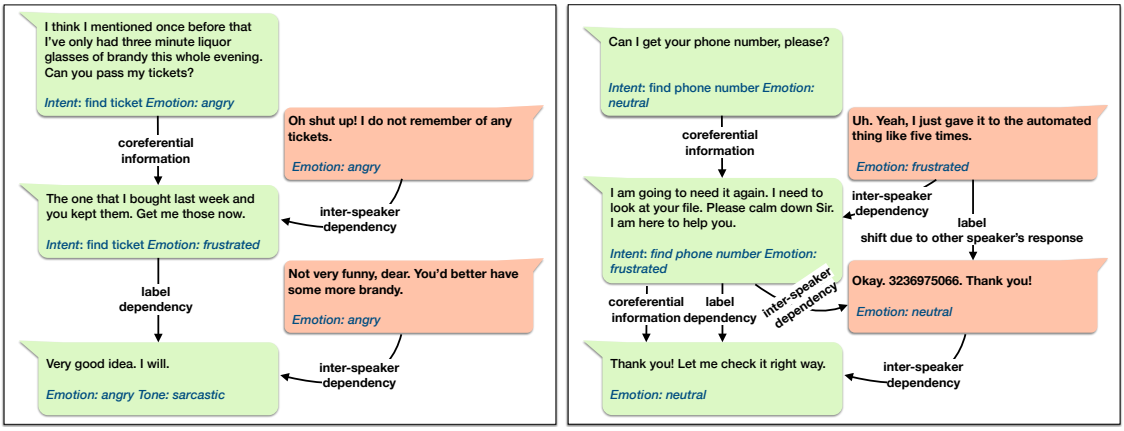

Figure 1: Role of Context in Utterance Level Dialogue Understanding.

Abstract

The recent abundance of conversational data on the Web and elsewhere calls for effective NLP systems for dialogue understanding. Complete utterance-level understanding often requires context understanding, partly defined by the nearby utterances and by the user intention and background. In recent years, a number of context-aware approaches have been proposed for various utterance-level dialogue understanding tasks. In this paper, we explore and quantify the role of context for different aspects of a dialogue, namely *emotion, dialogue act,* and *intent* identification, using state-of-the-art dialogue understanding methods as baselines. Specifically, we employ various perturbations to distort the context of a given utterance and study its impact on the different tasks and baselines. This provides us with insights into the fundamental context factors that have immediate implications on different aspects of a dialogue. Such insights may inspire more effective dialogue understanding models and provide support for future text generation approaches.

1 Introduction

Human-like conversational systems are a long-standing goal of Artificial Intelligence (AI). However, the development of such systems is not a trivial task, as we often participate in dialogues by relying on several contextual factors such as emotions, prior assumptions, intent, or personality traits. It is thus not surprising that the landscape of dialogue understanding research embraces several challenging tasks, such as, emotion recognition in conversations (ERC), dialogue intent classification, user-state representation, and others. These tasks are often performed at utterance level and can be conjoined together under the umbrella of *utterance-level dialogue understanding*. Due to the fast-growing research interest in dialogue understanding, several novel approaches have recently been proposed (Qin et al., 2020; Rashkin et al., 2019; Xing et al., 2020; Lian et al., 2019; Saha et al., 2020) to address the tasks by adopting speaker-specific and contextual modeling. However, to the best of our knowledge, the role of context has not been thoroughly explored across these tasks, partly due also to the lack of an unified framework across various utterance-level dialogue understanding tasks. In this work, we explore the role of context in utterance-level dialogue understanding. We use a contextual utterance-level dialogue understanding baseline (bcLSTM (Poria et al., 2017)) as

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1435–1449
August 1–6, 2021.

a strong baseline for the six dialogue-understanding tasks in four datasets. We propose several unique context probing strategies and experimental designs that test and measure: (1) speaker-specific context; (2) context order; (3) paraphrased context; (4) label shifts; (5) role of CRF in the sequence tagging of utterances in a dialogue. These strategies can be easily adapted for other tasks for similar purposes and provide insights into the development of new approaches to address these tasks.

Task Definition: Given a conversation along with speaker information of each constituent utterance, the utterance-level dialogue understanding task aims to identify the label of each utterance from a set of predefined labels that can be a set of emotions, dialogue acts, intents etc. Fig. 1 illustrates one such conversation between two people, where each utterance is labeled by emotion and intent. Formally, given the input sequence of N utterances $[(u_1, p_1), (u_2, p_2), \ldots, (u_N, p_N)]$, where each utterance $u_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,T}]$ consists of T words and spoken by party p_i, the task is to predict the label e_i of each utterance u_i. The classifier can make use of the conversational context in the process.

2 Models

We train our classification models in an end-to-end setup. We first extract utterance level features with a CNN module with pretrained GloVe vectors. The resulting features are non-contextual in nature as they are obtained from utterances without the surrounding context. We then classify the utterances with one of the following two models: i) Logistic Regression, or ii) bcLSTM. Among these, the Logistic Regression model is non-contextual in nature, whereas the bcLSTM is contextual. We expand on the feature extractor and the classifier in detail next.

2.1 Utterance Feature Extractor

Utterance level features are extracted using the following method:

GloVe CNN. A convolutional neural network (Kim, 2014) is used to extract features from the utterances of the conversation. We use a convolutional layer followed by max-pooling and a fully-connected layer to obtain the representation of the utterance. Each word in the utterances is initialized with 300d pretrained GloVe embeddings (Pennington et al., 2014). We pass these to convolutional filters of sizes 1, 2, and 3, each having 100 feature maps. The output of these filters are then max-pooled across all the words of an utterance. These are then concatenated and fed to a 100 dimensional fully-connected layer with ReLU activation (Nair and Hinton, 2010). The output after the activation form the final representation of the utterance.

2.2 Utterance Classifier

The representations obtained from the *Utterance Feature Extractor* are then classified using one of the following two methods:

Without Context Classifier. In this model, classification of an utterance is performed using a fully connected multi-layer perceptron layer. This classification setup is non-contextual in nature as there is no flow of information from the contextual utterances. We call this model **GloVe CNN**.

GloVe bcLSTM. The Bidirectional Contextual LSTM model (bcLSTM) (Poria et al., 2017) creates context-aware utterance representations by capturing the contextual content from the surrounding utterances using a Bi-directional LSTM (Hochreiter and Schmidhuber, 1997) network. bcLSTM is a strong contextual utterance-level dialogue understanding baseline, with consistent performance across all six dialogue-understanding tasks considered in this work. In our experiments, on an average bcLSTM is only 1% worse than the state of the art across the six tasks that we address in this work. As opposed to more complicated models (Majumder et al., 2019; Qin et al., 2020; Zhong et al., 2019), the simpler architecture of bcLSTM is devoid of complicated interactions amongst the contextual utterances, as attention. This enables easier interpretation of the effects of the perturbations of the context.

The feature representations extracted by the *Utterance Feature Extractor* serve as the input to the bcLSTM network. Finally, the context-aware utterance representations from the output of the bcLSTM are used for the label classification. The bcLSTM model is speaker independent as it does not model any speaker level dependency. In our implementation, we add a residual connection between the first and the output from the final layer to improve the network's stability. We call this model **GloVe bcLSTM**.

Why GloVe based CNN and LSTM-based Models: In this study, we consider GloVe CNN, GloVe bcLSTM and set up different scenarios to analyze them because these models are conceptually

much more straightforward than other state-of-the-art models such as DialogueRNN (Majumder et al., 2019) and DialogueGCN (Ghosal et al., 2019). For example, DialogueRNN also tracks the speaker states in addition to context. Thus, perturbations in the input would influence speaker modeling along with context modeling. This results in more complex deviations than bcLSTM, and are more difficult to analyze. Simple models are likely to be more interpretable. E.g., owing to DialogueRNN's complexity, we need to perform different levels of ablation studies to explain its behavior.

Furthermore, we use GloVe embeddings as recent transformer based models such as BERT (Devlin et al., 2018) is trained using the masked language model (MLM) objective that is already very powerful in modeling cross sentential context representation as demonstrated by other works (Liu et al., 2019; Lewis et al., 2019). Hence, to conduct a fair comparison between non-contextual and contextual models and further, for an easier apprehension on the role of contextual information in utterance-level dialogue understanding, we resort to the GloVe CNN and LSTM-based models. Additionally, as we perform a number of analysis studies, the GloVe based models were computationally much more efficient and faster to train and analyze.

3 Experimental Setup

3.1 Datasets

All the dialogue classification datasets that we consider in this work consists of two-party conversations in English language. We benchmark the models on the following datasets (see Table 1):

IEMOCAP (Busso et al., 2008) is a dataset of two person conversations among ten different unique speakers. The train set dialogues come from the first eight speakers, whereas the test set dialogues are from the last two. Each utterance is annotated with one of the following six emotions: *happy, sad, neutral, angry, excited,* and *frustrated*.

Dataset	# dialogues			# utterances		
	train	val	test	train	val	test
IEMOCAP	108	12	31	5163	647	1623
DailyDialog	11,118	1,000	1,000	87,179	8,069	7,740
MultiWOZ	8,438	1000	1,000	113,556	14,748	14,744
Persuasion	220	40	40	7902	1451	1511

Table 1: Statistics of splits and evaluation metrics used in different datasets. *Neutral** classes constitutes to 83% of the DailyDialog dataset. These are excluded when calculating the metrics in DailyDialog.

DailyDialog (Li et al., 2017) covers various topics about our daily life and follows the natural human communication approach. All utterances are labeled with both emotion categories and dialogue acts. The emotion can belong to one of the following seven labels: *anger, disgust, fear, joy, neutral, sadness*, and *surprise*. The dataset has over 83% *neutral* labels and these are excluded during Macro-F1 evaluation. In comparison, the dialogue act label distribution is relatively more balanced. The act labels can belong to the following four categories: *inform*, *question*, *directive*, and *commissive*.

MultiWOZ (Budzianowski et al., 2018) or Multi-Domain Wizard-of-Oz dataset is a fully-labeled collection of human-human conversations spanning over multiple domains and topics. The dataset has been created for task-oriented dialogue modelling and has 10,000 dialogues, which is atleast an order bigger than previously available task-oriented corpora. The dialogues are labelled with belief states and actions. It contains conversations between an user and a system from the following seven domains: restaurant, hotel, attraction, taxi, train, hospital and police. Here we focus on classifying the intent of the utterances from the user which belong to one of the following categories: *book restaurant*, *book train*, *find restaurant*, *find train*, *find attraction*, *find bus*, *find hospital*, *find hotel*, *find police*, *find taxi*, and *None*. The *None* utterances are not included in evaluation. Note that, utterances from the system side are not labelled and thus are not classified in our framework.

Persuasion For Good (Wang et al., 2019) dataset is a persuasive dialogue dataset where one participant aims to persuade the other participant to donate his/her earning using different persuasion strategies. The two participants are denoted as *Persuader* aka **ER** and *Persuadee* aka **EE** respectively. In this work, we formulate our problem to classify the utterances of Persuader and Persuadee separately using the full context of the conversation. This task can also be considered as a dialogue act classification task. The Persuader strategies are to be classified into the following eleven categories: *donation-information*, *logical-appeal*, *personal-story*, *foot-in-the-door*, *credibility-appeal*, *emotion-appeal*, *personal-related-inquiry*, *source-related-inquiry*, *self-modeling*, *task-related-inquiry*, and *non-strategy-acts*. The strategy can belong to one of the following thirteen categories for Persuadee,: *disagree-donation-more*, *ask-org-info*, *agree-donation*, *provide-donation-amount*,

Model	IEMOCAP	DailyDialog		MultiWOZ	Persuasion	
	Emotion	Emotion	Act	Intent	ER	EE
GloVe CNN	51.08	38.72	71.20	84.64	54.44	39.95
GloVe bcLSTM	61.90	**41.16**	**79.46**	**96.22**	**56.28**	**44.83**
w/o Inter-Speaker Dependency	**63.73**	39.99	74.50	95.05	53.24	40.63
w/o Intra-Speaker Dependency	56.45	35.93	78.69	95.75	52.23	38.93

Table 2: Classification performance in test data for the different tasks. Utterances from other speakers and the same speaker are absent respectively in the *w/o inter* and *w/o intra* settings. Scores are W-Avg F1 in IEMOCAP Emotion and MultiWOZ Intent; Macro F1 in the rest. All scores are average of 20 different runs. Test F1 scores are calculated at best validation F1 scores.

disagree-donation, *personal-related-inquiry*, *task-related-inquiry*, *ask-donation-procedure*, *negative-reaction-to-donation*, *positive-reaction-to-donation*, *ask-persuader-donation-intention*, *neutral-reaction-to-donation*, and *other-acts*.

3.2 Evaluation Metrics

In our experiments, we use Weighted average (W-Avg) F1 score in IEMOCAP emotion and MultiWOZ intent classification. For the other tasks – DailyDialog emotion, DailyDialog act, Persuader and Persuadee strategy classification – the label distribution is highly imbalanced, hence we report Macro F1 scores. In DailyDialog emotion classification, neutral labels are excluded (masked) while calculating the metrics. However, these utterances are still passed in the input of the different models.

4 Analysis

4.1 Speaker-specific Context Control

We first report the performance of the baseline GloVe CNN and GloVe bcLSTM model in the first two rows of Table 2. To further evaluate the intra- and inter-speaker dependence and relation across the different tasks in the GloVe bcLSTM model, we adopted two different settings as follows –

- **w/o Inter-Speaker Dependency:** when classifying a target utterance from speaker A, we drop the utterances of the speaker B from the context and vice versa.
- **w/o Intra-Speaker Dependency:** when classifying a target utterance from speaker A, we only keep utterances of the speaker B and drop all other utterances of speaker A from the context and vice versa.

Utterances of the Non-target Speaker are Important. The first setting coerces LSTM to only rely on the *target* speaker's (speaker of the target utterance) context in prediction. The results are reported in Table 2. As expected, performance drops are observed for all the datasets but IEMOCAP for emotion recognition, reinforcing the fact that the contextual utterances from the non-target speakers are important. Performance drop in DailyDialog dataset for act classification is noticeably the steepest. In the IEMOCAP dataset, we observe a pattern of the speakers maintaining the same emotion along a dialogue. This suggests that the speakers in the IEMOCAP dataset repeat the same emotion along consecutive utterances. Consequently, this induces a dataset bias. Hence, unlike the task of dialogue generation where the role of listener's utterance is key in generating speaker's response, we suspect in the case of emotion recognition in IEMOCAP dataset, removing other interlocutor's utterances from the context makes it easier and less confusing for the LSTM-based model to learn relevant contextual representations for the prediction. Contrary to this, although existing, repetitions of same or similar emotions in consecutive utterances of a speaker are less prevalent for emotion recognition in the DailyDialog dataset.

Utterances of the Target Speaker are *also* Important. 'w/o Intra-Speaker Dependency' scenario reported in Table 2 exhibits the importance of the utterances of the non-target speaker in the classification of the target utterance. In DailyDialog act and MultiWOZ intent classification, even when we remove the contextual utterances from the same speaker, the utterances from the non-target speaker provides key contextual information as evidenced by the performance in the 'w/o Intra-Speaker Dependency' setting. In those tasks, dropping the utterances of the non-target speaker results in more performance degradation as compared to the case when utterances from the target speaker are removed from the target utterance's context. This observation also supports the dialogue generation works (Zhou et al., 2017) that mainly consider previous utterances of the non-target speaker as the context for response generation. For emotion classification in DailyDialog and strategy classification in Persuasion For Good, the results obtained from 'w/o Intra-Speaker Dependency' setting are also relatively lesser compared to the baseline bcLSTM setting. This confirms the higher contextual salience of the target speaker's utterances over the non-target speaker's utterances for these particular tasks. In the case of the IEMOCAP emotion classification, removing the target speaker's utterances from the context causes a substantial

performance dip for the reasons stated earlier.

Interestingly, the 'w/o Inter-Speaker Dependency' setting in the DailyDialog dataset manifests two distinct trends for two different tasks – act classification and emotion recognition. While non-target speakers' utterances carry a little value for emotion recognition, they are extremely beneficial for act classification. This calls for task-specific context modeling techniques which should be the focus of the future works.

Key Takeaways of this Experiment. Although both target and non-target speakers' utterances are useful in several utterance-level tasks, we observe some divergent trends in some of the tasks in our experiments. Hence, we surmise that a task-agnostic unified context model may not be optimal in solving all the tasks. In the future, we should strive for task-specific contextual models as each task can have unique features that make it distinct from others. One can also think of multi-task architectures where two tasks can corroborate each other in improving the overall performance.

Logically, dropping contextual utterances in a dialogue leads to inconsistency in the context and consequently, it should degrade the performance of a model that relies on the context for inference. Hence, given an unmodified dialogue flow, an ideal contextual model is expected to refer to the right amount of contextual utterances relevant in inferring the label of a target utterance. In contrast, bcLSTM shows performance improvement for IEMOCAP emotion classification when utterances from the non-target speaker are dropped (refer to the 'w/o Inter-Speaker Dependency' row in Table 2). The performance does not change much for dialogue act and intent classification in the DailyDialog and MultiWOZ, respectively, when we drop utterances of the target speaker. These contrasting results indicate a potential drawback of the bcLSTM model in efficiently utilizing contextual utterances of both interlocutors in unmodified dialogues for the above mentioned tasks.

4.2 Classification in Shuffled Context

To analyze the importance of context, we shuffle the utterance order of a dialogue and try to classify the correct label from the shuffled sequence. For example, a dialogue having utterance sequence of $\{u_1, u_2, u_3, u_4, u_5\}$ is shuffled to $\{u_5, u_1, u_4, u_2, u_3\}$. This shuffling is carried out randomly, resulting in an utterance sequence whose order is different from the original sequence.

Context Shuffling Strategy			IEMOCAP	DailyDialog		MultiWOZ	Persuasion	
Train	Val	Test	Emotion	Emotion	Act	Intent	ER	EE
✗	✗	✗	**61.90**	**41.16**	**79.46**	**96.22**	**56.28**	**44.83**
✓	✓	✗	59.74	36.87	74.88	91.34	54.91	41.52
✗	✗	✓	57.63	34.58	66.81	67.91	50.69	37.17
✓	✓	✓	59.82	37.69	74.62	90.78	53.60	40.96

Table 3: Test performance of GloVe bcLSTM models in the different tasks for various shuffling strategies. In Train, Val, Test column ✓denotes shuffled context and ✗denotes unchanged context. Scores are W-Avg F1 in IEMOCAP Emotion and MultiWOZ Intent; Macro F1 in the rest. Test F1 scores are calculated at best validation F1 scores.

We design three such shuffling experiments: i) dialogues in train and validation sets are shuffled, test set is unchanged, ii) dialogues in train and validation sets are kept unchanged, but dialogues in test set are shuffled, iii) dialogues in train, validation and test sets are all shuffled.

We analyze these shuffling strategies in the GloVe bcLSTM model. In theory, the recurrent nature of the LSTM model allows it to be capable of modelling contextual information from the beginning of the sequence to the very end. However, when classifying an utterance, the most crucial contextual information comes from the neighbouring utterance. In an altered context, the model would find it difficult to predict the correct labels because the original neighbouring utterances may not be in immediate context after shuffling. This kind of perturbation would make the context modelling less efficient, and performance is likely to drop compared to their non-shuffled context counterparts. This is empirically shown in Table 3.

We observe that, whenever there is some shuffling in train, validation, or test set, the performance decreases a few points in all the datasets across all tasks and evaluation metrics. Notably, the performance drop is highest when the dialogues in train, validation sets are kept unchanged and dialogues in test set are shuffled. Note that, the result for this shuffling strategy (only test set is shuffled) in MultiWOZ stands at 67.91%, much lower than the original baseline of 96.22%. This is because, the test score of 67.91% is reported at the best validation score, even though we obtain better test scores at the initial epochs of training (around 78%).

Our reported results and observations are contradictory to the claims made by Sankar et al. (2019). According to Sankar et al. (2019), the shuffling of contextual utterances does not affect the response generation performance of a seq2seq model. There can be a number of reasons for these two contradicting observations: 1) first, the characteristics of utterance labels in a dialogue are different

#	Attack Method	Strategy Past	Strategy Future	Strategy Target	Window	IEMOCAP Emotion	DailyDialog Emotion	DailyDialog Act	MultiWOZ Intent	Persuasion ER	Persuasion EE
1	-	-	-	-	-	61.90	41.16	79.46	96.22	56.28	44.83
2	PA	✓	✗	✗	3	61.09	40.82	75.81	95.67	56.46	43.64
3	PA	✓	✗	✗	5	60.93	38.79	77.23	95.53	56.41	41.93
4	PA	✓	✗	✗	10	59.83	-	-	95.23	54.89	39.89
5	PA	✗	✓	✗	3	61.58	39.60	79.11	95.94	55.83	43.21
6	PA	✗	✓	✗	5	60.99	39.77	79.17	95.64	55.43	40.67
7	PA	✗	✓	✗	10	60.72	-	-	95.77	57.12	43.36
8	PA	✓	✓	✗	3	59.43	37.16	76.61	94.87	57.44	42.51
9	PA	✓	✓	✗	5	58.36	38.76	76.53	94.61	53.32	43.33
10	PA	✓	✓	✗	10	57.29	-	-	94.31	54.36	43.80
11	PA	✗	✗	✓	-	58.08	37.16	75.30	93.78	50.24	38.78
12	PA	✓	✓	✓	3	56.53	23.46	73.16	91.47	47.50	37.39
13	PA	✓	✓	✓	5	53.64	28.59	73.18	90.98	45.31	35.16
14	PA	✓	✓	✓	10	51.33	-	-	90.58	49.00	32.49

Table 4: Results for PA: *Paraphrasing-based Attack* in *utterance-based* GloVe bcLSTM model. In DailyDialog, we constrain the window size to 3 and 5 as there are an average of 8 utterances per dialogue in the dataset. Scores are W-Avg F1 in IEMOCAP Emotion and MultiWOZ Intent; Macro F1 in the rest.

from responses—responses are subjective and not unique, however labels are usually agreed upon by the observers to some degree—, 2) second, instead of reporting qualitative results, Sankar et al. (2019) only reported the perplexity score of their experiments. As stated in (Cai et al., 2019), perplexity and BLEU scores may not correctly represent the quality of the response generation.

4.3 Attacks with Context and Target Paraphrasing

Modern machine learning systems are often susceptible to attacks that slightly perturb the input without any drastic change in the semantics. Although prevalent in images, adversarial examples also exist in neural network-based NLP applications. In the context of NLP, crafting adversarial examples would require making character-, word-, or sentence-level modifications to the input text to trick the classifier into misclassification. Paraphrasing sentences is one such method to construct effective adversarial examples (Iyyer et al., 2018). We conduct several experiments to evaluate the sensitivity of utterance-level dialogue understanding systems to input paraphrasing. It should be noted that although task-specific adversarial strategies could be adopted, we chose to use a general set of attacking strategies in order to understand the behavior of the baseline across different tasks and datasets. This also facilitates a fair comparison among the tasks and whether there is a confounding factor that differentiates one task from another under the same attacking strategies.

Method. We use the following scheme to analyze this effect:

- The input utterances are modified at word level. For this modification, an average of 3 to 4 words are selected per utterance and masked. The pretrained RoBERTa model is then used to fill the masks with the most likely candidates. The utterance with substituted words form the new input. We call this method *Paraphrasing-based Attack* (PA).
- For each utterance (u_t) in a dialogue, we take a window of w immediate neighbouring utterances (context) on which the above modifications are performed. The window is selected as follows:
  - Only past w utterances: $u_{t-w}, .., u_{t-1}$
  - Only future w utterances: $u_{t+1}, .., u_{t+w}$
  - Past w and future w utterances: $u_{t-w}, .., u_{t-1}, u_{t+1}, .., u_{t+w}$
  - Past w, future w, and the target utterance: $u_{t-w}, .., u_{t-1}, u_t, u_{t+1}, .., u_{t+w}$
  - Only the target utterance: u_t

  In the last case, the window is empty. In other cases, we experiment with window size $w = 3, 5, 10$.

Model	IEMOCAP Emotion	DailyDialog Emotion	DailyDialog Act	MultiWOZ Intent	Persuasion ER	Persuasion EE
GloVe CNN	51.08	38.72	71.20	84.64	54.44	39.95
GloVe CNN PA	39.19(↓23.27)	23.82(↓39.64)	62.93(↓13.01)	70.34(↓16.89)	42.8(↓21.38)	33.59(↓15.91)
bcLSTM	61.90	41.16	79.46	96.22	56.28	44.83
bcLSTM PA	58.08(↓6.17)	37.16(↓9.71)	75.3(↓5.23)	93.78(↓2.53)	50.24(↓10.73)	38.78(↓13.49)

Table 5: Results for PA: *Paraphrasing-based Attack* in GloVe CNN model on target utterance and comparing it to bcLSTM results in Table 4. Scores are W-Avg F1 in IEMOCAP Emotion, MultiWOZ Intent; Macro F1 in the rest.

We train a GloVe bcLSTM and a GloVe CNN model with unadulterated train and validation data. During evaluation, however, the context and target are paraphrased as described before. The results of these experiments for bcLSTM and GloVe CNN are shown in Table 4 and Table 5, respectively.

Observations. We observe that the *Paraphrasing-based Attack* is quite effective in fooling the classifier in a number of tasks. The classification performance progressively deteriorates with larger window sizes.

In DailyDialog act classification, *Paraphrasing-based Attack* on only future utterances doesn't affect the results at all. The classification performance still remains very close to the original score of 79.46 %. We observe that there is a strong reliance on the label and content of past utterance in this task. For example, a *question* is likely to be followed by an *inform* or another *question* and much less likely to be followed by a *commissive* utterance. Unchanged past context thus results in

performance that is very close to the original setup. Attacking the past utterances combined with future and/or target utterances results in a relatively bigger performance drop. We also notice that the drop in performance is relatively much lesser than the other tasks except in MultiWOZ for intent classification. This is possibly because the act labels are mostly driven by the sentence type and hence unlikely to be affected from paraphrasing perturbations. For instance, around 30% of the act labels are of type *question*, and our attack strategy is almost guaranteed not to change an utterance with label *question* to something which might be classified as *inform*, *commissive*, or *directive*. Overall, we observe a consistent plunge in the performance when the target utterance is attacked by the *Paraphrasing-based Attack* method. For intent classification in MultiWOZ, utterances often have keywords which indicate the label (presence of *train* might indicate class label of *find train* or *book train*). In these cases, if the target utterance is not paraphrased, the model is still likely to predict the correct label. Finally, in Persuasion for Good, we observe that the attack method is slightly more effective in fooling the classifier for persuadee strategy classification.

In terms of window direction, we observe that perturbations in the past or future utterances result in a similar range of reduction in performances. One notable exception is act prediction in DailyDialog, where the model continues to perform near the original score of 79.46% irrespective of the attack in future utterances in the window.

Performance Comparison for Attacks in GloVe CNN and GloVe bcLSTM. We summarize the performance of GloVe CNN and GloVe bcLSTM models against *Paraphrasing-based Attack* in Table 5. For all the tasks, we observe a very significant drop in performance for GloVe CNN. For example, in emotion classification, the drop is around 23% and 40% for *Paraphrasing-based Attack* in IEMOCAP and DailyDialog respectively. However, for the same setting, the relative decrease in performance is only around 6% and 10% for bcLSTM. We observe the same trend in other tasks where it can be seen that the bcLSTM model is much more robust against the attack compared to the CNN model. This is because contextual models such as bcLSTM are harder to fool as the context carry key information regarding the semantics of the target and salient information can be inferred about the target using its' context. It is thus evident that even when the target utterance is corrupted, bcLSTM is capable of using contextual information to predict the label correctly, and subsequently the decline in performance is much lesser.

In principle, our findings in Table 5 can be related to how transformer-based pre-trained language models work. For example, in BERT (Devlin et al., 2018), the masked language modeling (MLM) and the next sentence prediction (NSP) objective forces the model to infer or predict the target using contextual information. Such contextual models are more powerful and robust because context information plays a crucial role in almost every natural language processing task. An objective similar to next sentence prediction in BERT or permutation language modeling in XLNET (Yang et al., 2019) can be used for conversation level pre-training to improve several downstream conversational tasks. Such approaches have been found to be useful in the past (Hazarika et al., 2019).

4.4 Performance for Label Shift

As discussed before, a few of our tasks of interest exhibit the label copying property which means consecutive utterances from the same speaker or different speakers often have the same or similar emotion, act, or intent label. The inter-speaker and intra-speaker label copying is especially prevalent in the IEMOCAP emotion task, the DailyDialog act task, and the MultiWOZ intent task. Contextual models such as bcLSTM make correct predictions when utterances display such kind of continuation of the same label. But what happens when there is a change of label? Does bcLSTM continue to perform at the same level or is it affected from the change? To understand this occurrence in more detail, we define this event as *Label Shift* and look at the following two different kind of shifts that could happen in the course of a dialogue:

- *Intra-Speaker Shift*: The label of the utterance is different from the label of the previous utterance from the same speaker.
- *Inter-Speaker Shift*: The label of the utterance is different from the label of the previous utterance from the non-target speaker.

In these two scenarios explained above, we are interested to see how bcLSTM performs at the utterances were the label shift takes place.

We report results for utterances in the test data that show *Intra-Speaker Shift* and *Inter-Speaker Shift* in Table 6. The *Inter-Speaker Shift* is not

Setup	IEMOCAP	Dailydialog		MultiWOZ	Persuasion	
	Emotion	Emotion	Act	Intent	ER	EE
Original	61.90	41.16	79.46	96.22	56.28	44.83
Intra-Speaker Shift	52.01 (13.2)	44.23 (1.0)	76.18 (2.9)	94.91 (1.6)	57.84 (6.9)	49.4 (4.7)
Inter-Speaker Shift	52.37 (22.0)	47.77 (1.3)	78.80 (4.9)	-	-	-

Table 6: Classification performance for utterances which exhibits *Label Shift* in test data. Numbers in parenthesis indicate the average count of the corresponding shifts per dialogue. There is no *Inter-Speaker Shift* in MultiWOZ or Persuasion for Good as we only classify user, persuader, or persuadee utterances. Scores are W-Avg F1 in IEMOCAP Emotion and MultiWOZ Intent; Macro F1 in the rest.

defined in MultiWOZ as we don't have intent labels for system utterances. We also don't report *Inter-Speaker Shift* results in Persuasion for Good as the persuader and persuadee strategy set is different.

The emotion labels in IEMOCAP display the largest extent of label copying. We also observe in Table 6 that label shifts occur with high frequency in IEMOCAP. These are the likely reasons why we observe significant number of errors for utterances with *Label Shift* for this task in Table 6. The performance for both *Intra-Speaker Shift* and *Inter-Speaker Shift* stands at around 52.0%, much lesser than the overall average of 61.9% in test data. Although not as strong as IEMOCAP, the intra-speaker label copying feature can also be seen in MultiWOZ intent and DailyDialog act labels. For these two tasks, we again observe a drop of performance when either *Intra-Speaker Shift* or *Inter-Speaker Shift* occurs. In contrast, the extent of transition is spread over a much larger combination of labels in DailyDialog emotion and Persuasion for Good. We observe that the results for utterances with *Label Shift* in those tasks are in fact better than the overall score. In DailyDialog emotion, the scores are 44.23% and 47.77%, which is an improvement over the original 41.16%. The scores of 57.84% and 49.4% in Persuasion for Good also stand over the scores of 56.28% and 44.83% in the original setup.

4.5 Sequence Tagging using Conditional Random Field (CRF)

On the surface, the task of utterance level dialogue understanding looks similar to sequence tagging. Are there any distinct label dependency and patterns across the tasks that are dataset agnostic and likely to be captured by CRF (Lafferty et al., 2001)? In the quest to answer this, we plug in three different CRF layers on top of the bcLSTM network.

Global-CRF. It is a linear chain CRF used on top of bcLSTM. In this setting, we do not consider speaker information.

It can be defined using the equations below:

$$P(Y|D) = \frac{1}{Z(D)} \prod_{i=1}^{n} \phi_T(y_{i-1}, y_i)\phi_E(y_i, u_i), \quad (1)$$

$$Z(D) = \sum_{y' \in \mathcal{Y}} \prod_{i=1}^{n} \phi_T(y'_{i-1}, y'_i)\phi_E(y'_i, u_i). \quad (2)$$

Global-CRF$_{\text{ext}}$. The linear-chain CRF is extended to include not only the transition potential from the previous label to the current label, but also from the prior-to-previous label. Concisely, the current label is predicated on the previous two labels. Therefore, the transition potential function ϕ_T takes one extra argument y_{i-2}. The advantage here is it also considers the previous label from the target speaker should utterance $i-2$ have come from the target speaker. This becomes useful in the tasks where the speakers tend to retain label from its last utterance. It can be defined using the equations below:

$$P(Y|D) = \frac{1}{Z(D)} \prod_{i=1}^{n} \phi_T(y_{i-2}, y_{i-1}, y_i)\phi_E(y_i, u_i), \quad (3)$$

$$Z(D) = \sum_{y' \in \mathcal{Y}} \prod_{i=1}^{n} \phi_T(y'_{i-2}, y'_{i-1}, y'_i)\phi_E(y'_i, u_i). \quad (4)$$

Speaker-CRF. In this setting, we use two distinct CRFs for the two speakers in a dialogue. Inter-speaker label dependency and transitions are not likely to be captured in this setting by the CRFs.

Negative Results. We report the results for CRF experiments in Table 7 and Table 8. Aside from the well-known sequence tagging tasks, such as, Named Entity Recognition (NER) and Part of Speech Tagging, CRF does not improve the performance of utterance-level dialogue understanding tasks. There could be multiple reasons as below:

1: A dialogue is governed by multiple variables or pragmatics, e.g., topic, personal goal, past experience, expressing opinions or presenting facts based on personal knowledge, and the role of the interlocutors. Hence, the response pattern can vary depending on these variables. The personality of the speakers add an extra layer of complexity to this which causes speakers to respond differently under the same circumstances. An identical utterance can be uttered with different emotions by two different speakers. CRF relies on surface label patterns which can vary with datasets. Due to this dynamic nature of dialogues and the presence of latent controlling variables, the label transition matrix of CRF does not learn any distinct

Methods	IEMOCAP	DailyDialog				
	Emotion	Emotion			Act	
	W-Avg F1	W-Avg F1	Micro F1	Macro F1	W-Avg F1	Macro F1
GloVe CNN	52.04	49.36	50.32	36.87	80.71	72.07
GloVe bcLSTM	61.74	52.77	53.85	39.27	**84.62**	79.12
w/o inter	**63.73**	52.39	52.86	**39.99**	81.32	74.50
w/o inter w/ speaker-CRF	62.94	52.47	54.04	39.77	81.19	74.12
w/ global-CRF	61.62	53.05	53.86	39.27	83.91	79.10
w/ global-CRF$_{ext}$	61.64	53.06	54.40	39.64	84.27	**79.25**
w/ speaker-CRF	62.21	**53.16**	**54.68**	**39.74**	84.15	79.20

Table 7: Classification performance in test data for IEMOCAP and DailyDialog using different CRF configurations. All scores are average of at least 10 different runs. Test F1 scores are calculated at best validation F1 scores.

Methods	MultiWOZ	Persuasion			
	Intent	Persuader (ER)		Persuadee (EE)	
	W-Avg F1	W-Avg F1	Macro F1	W-Avg F1	Macro F1
GloVe CNN	84.30	67.15	54.45	58.00	41.03
GloVe bcLSTM	**96.14**	**69.26**	55.27	61.18	42.19
w/o inter	95.05	67.81	53.24	59.44	40.63
w/o inter w/ speaker-CRF	94.11	68.13	54.45	58.93	40.16
w/ global-CRF/speaker-CRF	95.48	68.59	55.60	61.24	42.62
w/ global-CRF$_{ext}$	95.51	69.23	**56.80**	**61.89**	**43.68**

Table 8: Classification performance in test data for MultiWOZ and Persuasion for Good using different CRF configurations. All scores are average of atleast 10 different runs. Test F1 scores are calculated at best validation F1 scores. In MultiWOZ and Persuasion for Good, the global-CRF and speaker-CRF setting are identical as we only classify utterances coming from one of the speakers (user in MultiWOZ, persuader or persuadee in Persuasion for Good).

pattern that is complementary to what is learned by the feature extractor. **2**: Some of the datasets — IEMOCAP and MultiWOZ — contain distinct label-transition patterns between the same and distinct speakers e.g., the label copying feature in the IEMOCAP dataset where the same or similar emotions are repeated by the same or both the speakers. Similarly, in the MultWOZ dataset, the intent *book restaurant* to be frequently followed by the intent *find taxi*. **We believe the distinct label patterns in the IEMOCAP and MultiWoZ datasets is potentially one of the reasons why contextual models perform so well on these three datasets and tasks compared to the rest.** On these two daatsets, we expected bcLSTM w/ global-CRF to outperform vanilla bcLSTM. However, we do not observe any statistically significant improvement using bcLSTM w/ global-CRF over bcLSTM. We posit that the evident label-transition patterns that exist in these two datasets are straightforward to capture without a CRF. In fact, we also tried GloVe CNN with a CRF layer on it, and surprisingly the result was not significantly higher than that of GloVe CNN. This can be attributed to the absence of explicit contextual and label transition-based features in the CRF.

Results in IEMOCAP and Persuasion for Good Datasets. We observe a *minor* performance improvement in the IEMOCAP dataset using speaker-CRF for emotion recognition. This observation directly correlates to the experiment under "w/o Inter-Speaker Dependency" setting in Table 2 and can be largely attributed to the label copying feature in the IEMOCAP dataset as explained in the last paragraph. In "w/o Inter-Speaker Dependency" setting, contextual utterances of the speaker B are not utilized to classify utterances of speaker A vice versa. The results do not improve when we use speaker-level CRF on bcLSTM under the "w/o Inter-Speaker Dependency" setting. From these observations, we can conclude that CRF is not learning any distinct label dependency and transition patterns that are not learned by the feature extractor or bcLSTM alone.

Global-CRF$_{ext}$ shows significant performance improvement on the Persuasion for Good dataset. Some of the key controllable factors of the dialogues such as topics in this dataset are fixed and can be learned intrinsically by the classifier. The scope of the dialogues in this dataset is very limited as there are only two possible outcomes of the dialogues – *agree to donate*, and *disagree to donate*. Hence, there can be some label transition patterns learned by the Global-CRF$_{ext}$ using a larger label-context window in the transition potential.

5 Conclusion

In this paper, we explored the role of context for six utterance-level dialogue understanding tasks in four different datasets. Using a strong contextual baseline system (bcLSTM), we gained insights into the behavior of such contextualized models in the presence of various context perturbations. Such probes have bolstered many interesting intuitions about utterance-level dialogue understanding—the role of label dependency and future utterances; the role of speaker-specific contextual modelling; and the robustness of contextual models as opposed to their non-contextual counterparts against adversarial probes. We believe that these probing strategies can be straightforwardly adapted to other context-reliant tasks. The implementation pertaining to this work is available at https://github.com/declare-lab/dialogue-understanding.

Acknowledgements

This research is supported by A*STAR under its RIE 2020 Advanced Manufacturing and Engineering programmatic grant, Award No.– A19E2b0098.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1219–1228.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2019. Emotion recognition in conversations with transfer learning from generative conversation modeling. *arXiv preprint arXiv:1910.04980*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885.

Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. 2017. Residual lstm: Design of a deep recurrent architecture for distant speech recognition. *arXiv preprint arXiv:1701.03360*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP 2014*, pages 1746–1751.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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 the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang. 2019. Domain adversarial learning for emotion recognition. *arXiv preprint arXiv:1910.13807*.

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

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proceedings of the Association for Computatinal Linguistics*.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv preprint arXiv:1906.01603*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. 2006. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2):165–183.

Songlong Xing, Sijie Mai, and Haifeng Hu. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.

A Label Transitions.

To check whether there lies any patterns in the label sequences of the datasets, in Fig. 2 and 3, we plot frequency of the label pairs (x, y) where x and y are the labels of $U_{s_{t-1},t-1}$ and $U_{s_t,t}$ respectively. Figure Fig. 2 explains inter-speaker label transition and Fig. 3 illustrates the intra-speaker label transition. Both these plots reveal the same emotion labels appearing in the consecutive utterances with high frequency in the IEMOCAP dataset. This induces label dependencies and consistencies and can be called as the **label copying feature** of the dataset. From our empirical analysis in Section 4, we confirm this property of the IEMOCAP dataset. Although not as strong as IEMOCAP, the intra-speaker label copying feature is also prevalent in the MultiWOZ and DailyDialog (Act) dataset (refer to Fig. 2). Moreover, we observe interesting patterns in DailyDialog (Act). A directive utterance is commonly followed by a commissive utterance. This indicates that utterances with acts such as request and instruct (directive label) are followed by accepting/rejecting the request or order (commissive label). We also notice that an utterance with the act of questioning is commonly followed by the utterances with the act of answering (which is quite natural). Fig. 2 also corroborates the high frequent joint appearance of similar emotions in both speaker's utterances e.g., negative emotions — *anger, frustration, sad* expressed by one speaker is replied with a similar negative emotion by the other speaker. Interestingly, the DailyDialog dataset for emotion classification does not elicit any such patterns. We can attribute this to the scripted utterances present in the IEMOCAP that has specifically been designed to invoke more emotional content to the utterances. On the other hand, the DailyDialog dataset comprises naturalistic utterances that are more dynamic in nature as they depend on interlocutors' personality. In both IEMOCAP and DailyDialog datasets, the repetitions of the same emotions can be found in consecutive utterances of a speaker. The repetition of the same or similar emotions for a speaker is frequent and often forms long chains in IEMOCAP. However, such repetitions are much less prevalent in DailyDialog. Readers are referred to Fig. 3 for a clearer view. **This two different types of datasets used in this work is purposefully crafted in order to study dataset-specific nuances to attempt the same task.** In DailyDialog, approximately 80% of utterances are labeled as *no-emotion* (see Fig. 4) which poses a difficult challenge to perform emotion classification. These two datasets also differ from each other in the average dialogue length. While the average number of utterances per dialogue in the IEMOCAP dataset is more than 50, the average number of utterances per dialogue in the DailyDialog dataset is just 8 which is much shorter.

Among other semantically plausible label transitions, we can see in Fig. 3, the intent *book restaurant* to be frequently followed by the intent *find taxi* in the MultiWOZ dataset. **We believe this is potentially one of the reasons why contextual models perform so well on these three datasets and tasks** compared to the rest which we discuss in the subsequent sections. Further, label dependency and consistency can aid filtering likely labels given the prior labels. Notably, such patterns are not visible in the other datasets. Hence, one can use Conditional Random Field (CRF) to find any hidden label patterns and dependencies.

B Utterance Classifier

cLSTM. Similar to bcLSTM but without the bidirectionality in the LSTM, this model is intended to ignore the presence of future utterances while classifying an utterance U_t.

DialogueRNN. (Majumder et al., 2019) is a recurrent network based model for emotion recognition in conversations. It uses two GRUs to track individual speaker states and global context during the conversation. Further, another GRU is employed to track emotion state through the conversation. In this work, we consider the emotion state to be a general state which can be used for utterance level classification (i.e., not limited to only emotion classification). Similar to the bcLSTM model, the features extracted by the *Utterance Feature Extractor* is the input to the DialogueRNN network. DialogueRNN aims to model inter-speaker relations and it can be applied on multiparty datasets.

cLSTM, bcLSTM and DialogueRNN with Residual Connections. Deep neural networks can often have difficulties in information prorogation. Multi-layered RNN-like in particulars often succumb to vanishing gradient problems while modeling long range sequences. Residual connections or skip connections (He et al., 2016) are an intuitive way to tackle this problem by improving information propagation and gradient flow. Inspired by the early works in residual LSTM (Wu

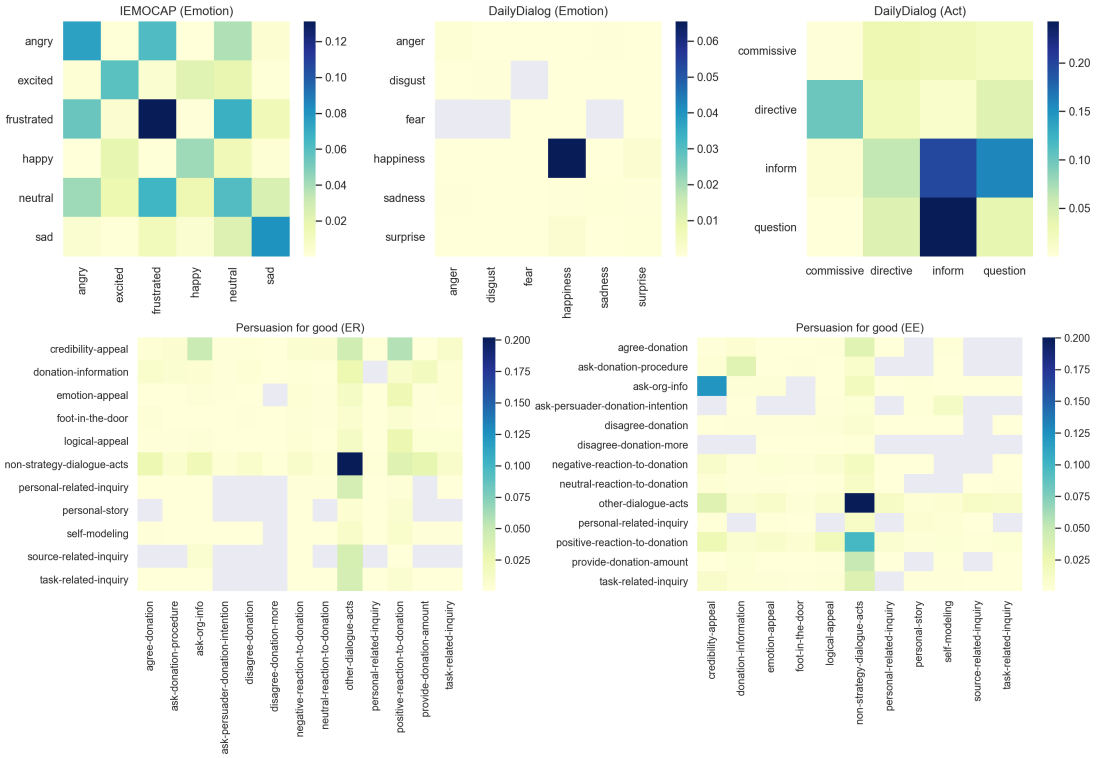

Figure 2: The heatmap of inter-speaker label transition statistics in the datasets. The color bar represents normalized number of inter-speaker transitions such that elements of each matrix add up to 1. Inter-speaker transitions are not defined in MultiWOZ as system side utterances are not labeled. Note: For the DailyDialog dataset, we ignore the *neutral* emotion in this figure.

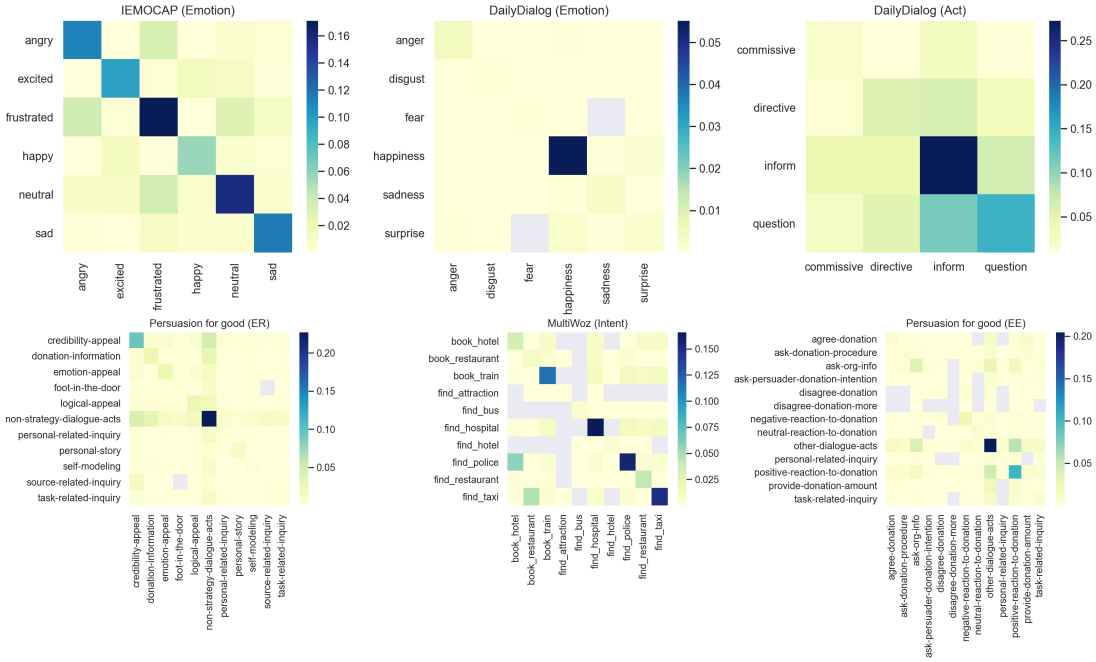

Figure 3: The heatmap of intra-speaker label transition statistics in the datasets. The color bar represents normalized number of intra-speaker transitions such that elements of each matrix add up to 1. Note: For the DailyDialog dataset, we ignore the *neutral* emotion in this figure.

et al., 2006; Kim et al., 2017), in our recurrent contextual models - bcLSTM and DialogueRNN we adopt a simple strategy to introduce a residual connection. For each utterance, a residual connection is formed between the output of the feature extractor and the output of the bcLSTM/DialogueRNN module. These two vectors are added and the final classification is performed from the resultant

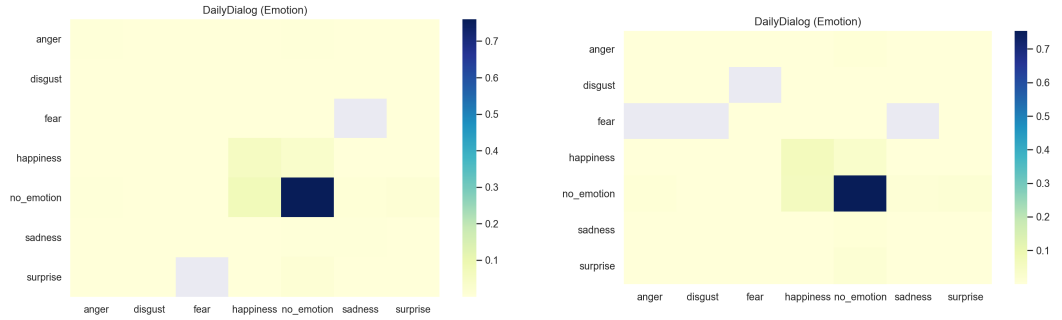

Figure 4: The heatmap of intra-speaker (left) and inter-speaker (right) label transition statistics in the DailyDialog dataset including *neutral* emotion. The color bar represents normalized number of inter-speaker and intra-speaker transitions such that elements of each matrix add up to 1.

vector.

B.1 Results

Methods	IEMOCAP	DailyDialog				
	Emotion	Emotion			Act	
	W-Avg F1	W-Avg F1	Micro F1	Macro F1	W-Avg F1	Macro F1
GloVe CNN	52.04	49.36	50.32	36.87	80.71	72.07
GloVe cLSTM	59.10	52.56	53.67	38.14	83.90	78.89
w/o Residual	55.07	52.56	53.26	38.12	84.06	78.54
GloVe bcLSTM	61.74	52.77	53.85	39.27	84.62	79.12
w/o Residual	58.32	54.74	**56.32**	39.24	84.10	78.98
GloVe DialogueRNN	**62.57**	**55.18**	55.95	**41.80**	**84.71**	**79.60**
w/o Residual	61.32	54.50	55.29	40.05	83.98	79.16
RoBERTa LogReg	54.12	52.63	52.42	40.02	82.55	75.62
RoBERTa bcLSTM	62.72	56.05	56.77	43.26	85.17	82.16
w/o Residual	62.86	55.92	57.32	43.03	**86.35**	80.69
RoBERTa DialogueRNN	**64.12**	**59.07**	**59.50**	**45.19**	86.31	**82.20**
w/o Residual	63.96	57.57	57.76	44.25	86.28	82.08

Table 9: Classification performance in test data for emotion prediction in IEMOCAP, emotion prediction in DailyDialog, and act prediction in DailyDialog. Scores of the Glove-based models are reported after averaging 20 different runs. RoBERTa-based models were run 5 times and we report the average scores. Test F1 scores are calculated at best validation F1 scores.

Methods	MultiWOZ	Persuasion			
	Intent	Persuader (ER)		Persuadee (EE)	
	W-Avg F1	W-Avg F1	Macro F1	W-Avg F1	Macro F1
GloVe CNN	84.30	67.15	54.33	58.00	41.03
GloVe cLSTM	95.03	68.75	54.36	59.46	41.62
w/o Residual	95.12	64.62	49.08	54.87	36.36
GloVe bcLSTM	96.14	**69.26**	55.27	**61.18**	**42.19**
w/o Residual	96.21	67.20	52.75	55.02	37.72
GloVe DialogueRNN	**96.32**	68.96	**56.29**	61.11	42.18
w/o Residual	96.08	68.77	54.20	58.72	39.06
RoBERTa LogReg	85.70	71.98	60.36	63.45	**51.74**
RoBERTa bcLSTM	95.46	71.85	61.05	64.14	50.11
w/o Residual	**95.61**	71.06	58.72	62.73	44.74
RoBERTa DialogueRNN	**95.61**	**72.91**	**62.03**	**64.33**	49.22
w/o Residual	95.29	72.45	60.49	64.21	49.71

Table 10: Classification performance in test data for intent prediction in MultiWOZ, persuader and persuadee strategy prediction in Persuasion for Good. Scores of the Glove-based models are reported after averaging 20 different runs. RoBERTa-based models were run 5 times and we report the average scores. Test F1 scores are calculated at best validation F1 scores.

We report results for IEMOCAP, DailyDialog dataset in Table 9 and MultiWOZ, Persuasion for Good dataset in Table 10. We ran each experiment multiple times and report the average test scores based on the best validation scores.

We observe that there is a general trend of improvement in performance when moving to the RoBERTa based feature extractor from the GloVe CNN feature extractor except in the intent prediction task in MultiWOZ dataset. As the RoBERTa model has been pre-trained on a large amount of textual data and has considerably more parameters, this improvement is expected. The results could possibly be improved even more if a RoBERTa-Large model is used instead of the RoBERTa-Base model that we use in this work.

We also observe that contextual models — bcLSTM and DialogueRNN perform much better than the non-contextual Logistic Regression models in most cases. Context information is crucial for emotion, act, and intent classification and models such as bcLSTM or DialogueRNN are some of the most prominent methods to model the contextual dependency between utterances and their labels. In IEMOCAP, DailyDialog and MultiWOZ there is a sharp improvement in performance in contextual models compared to the non-contextual models. However, for the strategy classification task in *Persuasion for Good dataset*, the improvement in contextual models is relatively lesser. Notably, for Persuadee classification, the RoBERTa non-contextual model achieves the best result, outperforming the contextual models. Without the presence of residual connections, the *GloVe cLSTM* and *GloVe bcLSTM* baselines perform poorly than the non-contextual *GloVe CNN* baseline in the *Persuasion for Good dataset*. This beckons the need for better contextual models for this dataset. To analyze the results of the different models we look at the following aspects:

Importance of the Residual Connections in the Models. It is also to be noted that the introduction of the residual connections generally improves

the performance of the contextual models. We obtain better performance and improved stability during training for most of the models with residual connections. In particular, residual connections are mostly effective in IEMOCAP and Persuasion for Good datasets that comprise long dialogues. Residual connections are used in deep networks to aid information propagation and tackle vanishing gradient problems (Wu et al., 2006; Kim et al., 2017) in RNNs by improving gradient flow. As multi-layered RNN-like architectures often find it difficult to model long-range dependencies in a sequence due to vanishing gradient problems (Pascanu et al., 2013), we conjecture, that could be one of the reasons why we see a great performance boost with residual connections by helping propagate key information form the CNN layers to the output of LSTM layers that might be lost due to the long deep sequence modeling in the LSTM layer. Residual connections also help in combating vanishing gradient issues by improving gradient flow. Unlike IEMOCAP and Persuasion for Good, in DailyDialog and MultiWOZ datasets, the improvement in performance caused by the residual connections is only little which can be attributed to the relatively shorter dialogues present in these two datasets.

Variance in the Results. As deep learning models tend to yield varying results across multiple training runs, we trained each model multiple times and report the average score in Table 9 and Table 10. In general, we observed that the RoBERTa-based models show lesser variance compared to the GloVe-based models.

Variance in the Glove-based models: The observed variance is higher for emotion classification in IEMOCAP and DailyDialog as compared to act and intent classification in DailyDialog and MultiWOZ, respectively. Both baseline models – Glove CNN and bcLSTM show standard deviation of about 1.28% in the IEMOCAP dataset across different runs. In the Persuasion for Good dataset, for both persuader's and persuadee's act classification tasks, the deviation remains around 1.6% when we consider the Macro-F1 metric. However, for the Weighted-F1 metric, the performance is relatively stable as upon accumulating multiple runs the standard deviation is about 0.99% across the baselines. A similar trend is also prevalent in the DailyDialog dataset for emotion classification. In this task, the baselines – Glove CNN and bcLSTM show standard deviation of about 1.19% when Weighted-F1 and Micro-f1 are considered. According to Macro-F1 metric, however, these baselines are exposed to relatively higher standard deviation of 2.88%. This is likely to be a consequence of severe label imbalance in the dataset, that is having 80% neutral utterances. We have observed that a majority of these neutral samples do not exhibit neutral emotion. Therefore, this poor labeling quality may have precipitated this large variance in the results. On the other hand, the baseline models perform consistently in the intent and act classification tasks in MultiWOZ and DailyDialog datasets respectively showing standard deviation of around 0.55% across different runs. When comparing among the baselines, we found higher variances in the results obtained with the Glove CNN than the bcLSTM.

One possible reason behind the variances in the results of the GloVe-based models could be the end-to-end training setup that renders the model deeper. The original bcLSTM and DialogueRNN model employed a two-stage training method where the utterance feature extractor is first pretrained and then kept unchanged during the contextual model training. This setting may make those original models more stable. Similarly, we think, in our end-to-end setup, a more sophisticated training regime could result in a lesser variance of the results. For example, the utterance feature extractor could be trained only for the first few epochs and then kept frozen during subsequent epochs of the training. Due to this high variance in the end-to-end Glove-based models, the future works on these datasets and tasks which employ this setting should report the average results of multiple runs for a fair comparison of the models.

Variance in the RoBERTa-based models: The RoBERTa based models show much lesser variance in performance across different runs. In particular, the standard deviations in the results of Roberta-based bcLSTM are 0.57 on the IEMOCAP, 0.08, and 0.48 in the DailyDialog for emotion and act classification tasks, respectively, 0.07 in the MultiWoz dataset, 0.9 and 1.04 in the Persuasion for Good dataset for persuader's and persuadee's act classification tasks respectively. RoBERTa-based DialogueRNN shows a similar trend. We surmise that this is the case because the feature extractor's weights are initialized from a pretrained checkpoint. Thus, the feature extractor already provides meaningful features from the beginning of training and is not required to be trained from scratch, resulting in greater stability in the performance.