A Practical Dialogue-Act-Driven Conversation Model for Multi-Turn Response Selection

Harshit Kumar
IBM Research,
Delhi, India
harshitk@in.ibm.com

Arvind Agarwal
IBM Research,
Delhi, India
arvagarw@in.ibm.com

Sachindra Joshi
IBM Research,
Delhi, India
jsachind@in.ibm.com

Abstract

Dialogue Acts play an important role in conversation modeling. Research has shown the utility of dialogue acts for the response selection task, however, the underlying assumption is that the dialogue acts are readily available, which is impractical, as dialogue acts are rarely available for new conversations. This paper proposes an end-to-end multi-task model for conversation modeling, which is optimized for two tasks, dialogue act prediction and response selection, with the latter being the task of interest. It proposes a novel way of combining the predicted dialogue acts of context and response with the context (previous utterances) and response (follow-up utterance) in a crossway fashion, such that, it achieves at par performance for the response selection task compared to the model that uses actual dialogue acts. Through experiments on two well known datasets, we demonstrate that the multi-task model not only improves the accuracy of the dialogue act prediction task but also improves the MRR for the response selection task. Also, the cross-stitching of dialogue acts of context and response with the context and response is better than using either one of them individually.

1 Introduction

Response selection remains at the core of conversation modeling, with the objective of selecting an appropriate response utterance from a set of candidate utterances, for a given conversation history consisting of previous utterances (context). Decades of research in this task includes traditional methods such as (Kitano, 1991; Ritter et al., 2011) and recent deep learning based methods (Ji et al., 2014; Chaudhuri et al., 2018; Xu et al., 2018; Chen et al., 2017; Song et al., 2018; Lowe et al., 2015; Zhao et al., 2018; Wen et al., 2016). Underlying these methods, a fundamental need is to capture the semantics of the context and use it for selecting the appropriate response. While the context provides essential clues as to what could be a follow-up response, research (Kumar et al., 2018) has further shown that any additional information available in the form of dialogue acts can also be helpful for response selection. Such information when used along with the context improves the performance of the response selection task. However, the above method assumes that dialogue acts are available at the time of response selection, which is rarely the case —as dialogue acts are usually not available for new conversations in a live setting— thus making them impractical for practitioners. In this paper, we propose a novel model that bridges this gap between theory and practice. In other words, our proposed model leverages the dialogue acts for response selection, as well as is practical.

In the literature, researchers (Kumar et al., 2018; Xu et al., 2018; Zhao et al., 2017) have proposed deep learning models that use actual dialogue acts in conversation modeling. While actual dialogue acts help in response selection, a natural question is, can we build a system that eliminates the dependency on *actual* dialogue acts at the time of response selection, and rather predict them as an integral part of the model? Second, and a more important question is: Is such a system going to be helpful in response selection, because the dialogue acts predictions will have some error in it, i.e., the underlying prediction model would not be 100% accurate in its predictions? And, if the answer to the second question is positive, then what is the gap - in terms of performance - between the proposed system that uses predicted dialogue acts and the system that uses actual dialogue acts? In this paper, we answer all of the above questions: our proposed model is a multi-task model that has dialogue acts prediction as an integral part of it,

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1980–1989,
Hong Kong, China, November 3–7, 2019.

i.e., it does not need the actual dialogue acts to select an appropriate response, rather it predicts the dialogue acts and use them for response selection. Furthermore, our model is by design robust to the errors in dialogue act prediction; our novel way of combining dialogue acts of context and response, is able to compensate for the errors in dialogue act predictions, and performs on par with the model that uses actual dialogue acts.

The main contributions of this paper are as follows:

- We model the task of response selection as a multi-task learning problem, with the objective of performing two tasks in a single end-to-end model: first, learn to predict the dialogue acts of utterances (context and response), and second, use the previous utterances (context) and the predicted dialogue acts of both the context and the response to select a response from a given set of candidate responses.

- While modeling the response selection conditioned on the dialogue acts of the context helps (Zhao et al., 2017), an important contribution is the additional utility of the dialogue act of the response. Our simple yet novel way of combining the dialogue act representations of the context and response with the utterance representations of the context and response promotes cross similarities, and thereby bring in ensemble characteristics in the model. That is, the ensemble model outperforms all other non-ensemble models, and is robust to the errors made by any underlying components of the ensemble.

- We evaluate the proposed model on two dialogue datasets, DailyDialog(Li et al., 2017) and Switchboard Dialogue Act Corpus (SwDA (Jurafsky, 1997)), and show that having dialogue act prediction as an integral part of the model improves the performance of the response selection consistently across both datasets. An important observation is the significant performance boost obtained from the proposed Crossway model (ensemble-model); that is, it not only improves the MRR for the response selection task but also improves the accuracy of the dialogue act prediction task.

2 Approach

This section details our approach, i.e. an end-to-end multi-task model for response selection (task1) using predicted dialogue acts (task2) of context and response. For response selection, there are two frameworks that are popular in the literature; one is generative and the other is discriminative. In the generative framework, sequence-to-sequence kind of model is used. It is trained to *generate* an appropriate response given a context. On the other hand, a discriminative model is trained such that among a set of K candidate responses, the correct response has the highest similarity with the context. Since discriminative model is superior than their generative counterpart(Liu et al., 2016), we use discriminative model as a base model.

Before proceeding, we first define the mathematical notation that we use throughout this work. Let $\mathcal{D} = (C^1, C^2, \ldots C^N)$ be a set of N conversations, with $(Y^1, Y^2, \ldots Y^N)$ be their corresponding actual DAs. Each conversation C^i is a sequence of R_i utterances $C^i = (u^i_1, u^i_2, \ldots u^i_{R_i})$ with $Y^i = (y^i_1, y^i_2, \ldots y^i_{R_i})$ being the corresponding actual DAs. For the notational simplicity, we shall ignore the conversation superscript i which should be clear from the context. For each utterance u_j in each conversation, we have an associated DA label $y_j \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all possible DAs. Each utterance u_j is a sequence of S_j words stringed together, i.e., $u_j = (w_{j1}, w_{j2}, \ldots w_{jS_j})$. In our problem setting, the first R_{i-1} utterances in a conversation C^i form the context, i.e. $context^i = (u^i_1, u^i_2, \ldots u^i_{R_i-1})$; and, the last utterance, $u^i_{R_i}$ is the true response. An illustration of the multi-task model for dialogue act prediction and response selection is illustrated in Figure 1.

Our approach of joint modeling of dialogue act prediction task and response selection task, share a common encoder which encodes the conversation context and response. These representations are then used to predict the dialogue acts and to find the right response from a set of candidate responses. In the following subsections, we provide details of this shared encoder, dialogue act prediction modeling, and response selection modeling.

2.1 Shared Context-Response Encoder

In each conversation, the whole sequence of utterances that constitutes a conversation can be con-

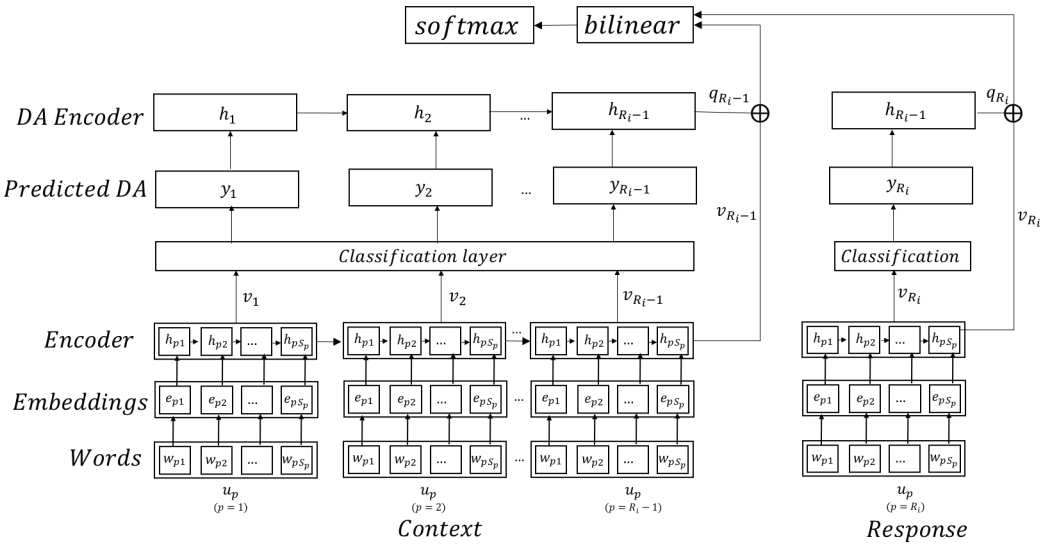

Figure 1: Architecture diagram of the proposed multi-task Crossway model for response selection and dialogue prediction

sidered as a single very long chain of words. This is input to an RNN encoder to obtain a single *unified* representation of the context (and response), and a representation of each utterance in the context (response). Given a conversation consisting of R_i utterances, with each utterance u_j consisting of S_j words, the sequence of operation used in encoder is as follows:

$$e_{jk} = f^1_{embed}(w_{jk}) \quad \forall j \in 1, \dots R_i, \forall k \in 1, \dots S_j$$
$$h_{jk} = f^1_{rnn}(h_{jk-1}, e_{jk}) \quad \forall j \in 1, \dots R_i, \forall k \in 1, \dots S_j \tag{1}$$

where, f^1_{embed} represents the embedding layer, whereas f^1_{rnn} is the encoder (RNN). The representation of each utterance u_j, denoted by v_j, can be obtained by combining the representations of its constituent words. We take the representation of the last time-step of the encoder as the representation of the entire utterance, i.e. $v_j = h_{jS_j}$. This is because the final time-step contains the context of all the words preceding it, and serves as a good approximation to the representation of the entire utterance. Thus the shared encoder finally gives us the representations of each utterance, i.e. $v_1, v_2, \dots v_{R_i-1}$, corresponding to the $context^i$ consisting of utterances $u_1, u_2, \dots u_{R_i-1}$, with v_{R_i-1} being the representation of the entire context. Since the encoder is shared between context and response, it is also used to get the representation v_{R_i} corresponding to response utterance u_{R_i}.

2.2 Dialogue Act Prediction Model

Dialogue act prediction (Task1) is a multi-class classification problem, where the goal is to assign a dialogue act to each utterance in a conversation. Following the recent advances in sequence prediction task (Kumar et al., 2017), the dialogue act prediction model is built on top of an RNN network, where each utterance's representation is first obtained using an RNN Encoder (2.1), which is then input to a classification layer to predict the appropriate dialogue act of that utterance. Given the representation of an utterance obtained using shared context-response encoder, the probability of predicting the dialogue act of the utterance u^i_k can be written as:

$$p(y|v_k) = \frac{\exp(-W_y^T v_k)}{\sum_{y' \in \mathcal{Y}} \exp(-W_{y'}^T v_k)} \tag{2}$$

where v_k is the encoded representation of utterance u_k. W_y is the weight vector associated with the class y. The network is optimized to maximize the probability of the gold-standard (actual) dialogue act. For the dialogue acts associated with the utterances in the context, the loss function can be written as following:

$$\mathcal{L}_c = \sum_{C^i \in \mathcal{D}} \sum_{u^i_k \in C_i \setminus u^i_{R_i}} -\log p(y^i_k | v^i_k) \tag{3}$$

where y^i_k is the actual dialogue act of the utterance u^i_k. $\mathcal{L}_c$ is the loss (i.e., negative log likelihood) computed from the prediction task of the context. We can compute a similar loss for the response as

following:

$$\mathcal{L}_r = \sum_{C^i \in D} -\log p(y^i_{R_i} | v^i_{R_i}) \quad (4)$$

where $v^i_{R_i}$ is representation of the response utterance $u^i_{R_i}$ obtained from the encoder, and $y^i_{R_i}$ is the corresponding actual dialogue act.

2.3 Dialogue-Act Aware Response Selection

The goal of the second task is to select the true candidate response from a set of candidate responses for a given context. This model consists of two modules, the first module is a Dialogue-Act Encoder which gives us two representations: a compositional representation for the sequence of dialogue acts associated with the context; and, a representation for the response dialogue act. The second module is a crossway response selection module which uses both dialogue act representations to select the right response from a set of candidate responses. This module combines the dialogue act representation and utterance representation of context utterances and response in a cross-stitched way using a Siamese network for the response selection task.

2.3.1 Dialogue-Act Encoder Module

In conversation modelling, dialogue acts are treated as an additional sequence of signals that can aid in the learning process. Dialog-Act encoder (DA-encoder), which is based on the same principle as the RNN encoder, takes the sequence of dialogue acts and returns a representation of that sequence. The input to the DA-encoder are one hot encodings of the dialogue acts, which are then passed through an embedding layer (f^2_{embed}) to learn DA embeddings. These DA embeddings are sent to an RNN (f^2_{rnn}) to learn a representation of the entire DA sequence. For a given sequence of DA of length K, the sequence of operations for the DA-encoder is as follows:

$$\begin{aligned} e_k &= f^2_{embed}(y_k) \quad \forall k \in 1, 2, \dots K \\ h_k &= f^2_{rnn}(h_{k-1}, e_k) \quad \forall k \in 1, 2, \dots K \\ q_K &= h_K \end{aligned} \quad (5)$$

where, q_K is the final representation of the dialogue act sequence.

2.3.2 Crossway Response Selection Module

The Crossway Response Selection Module uses the shared context-response encoder to get the representations of the utterances in a context, i.e. v_{R_i-1}, and the response, i.e. v_{R_i}; and DA encoder to get the representation of the DA sequence associated with the context, i.e q_{R_i-1} and response, i.e. q_{R_i}. A typical discriminative model, or in particular Siamese model, consists of two encoders, one encoder encoding the context, while another encoding the response utterance. These two representations are passed to a final layer that computes the probability of candidate being a valid response given the context. In the previous response selection models that use dialogue acts, authors have only used the dialogue act representations of the context and not of the response. We use all four representations in a Crossway fashion, i.e. v_{R_i-1}, v_{R_i}, q_{R_i-1} and q_{R_i}. As we shall see in the experiments, using these four representation adds robustness to the Crossway model. The two representations corresponding to context and its DA sequence, v_{R_i-1} and q_{R_i-1}, are concatenated together to obtain a compositional representation of the context. Similarly, the two representations of the response and its associated dialogue act, v_{R_i} and q_{R_i} are concatenated together to obtain a compositional representation of the response utterance. The probability of the association of these representations can be computed using a bilinear function as following:

$$\begin{aligned} d_{R_i-1} &= [v_{R_i-1}, q_{R_i-1}] \\ d_{R_i} &= [v_{R_i}, q_{R_i}] \\ p(s|d_{R_i-1}, d_{R_i}) &= \sigma(d^T_{R_i-1} \cdot A \cdot d_{R_i} + b) \end{aligned} \quad (6)$$

where, the bias b and matrix A are learned model parameters. The model is trained by minimizing the cross-entropy of all labeled conversations including positive and negative examples. Let $\mathcal{D}^-$ be the variation of $\mathcal{D}$ where response utterance $u^i_{R_i}$ is replaced with some random utterance in order to create negative examples. Given the set of positive and negative conversation sets, the loss is computed as follows:

$$\begin{aligned} \mathcal{L}_s = &- \sum_{C^i \in \mathcal{D}} \log p(s^i = 1 | d_{R_i-1}, d_{R_i}) \\ &- \sum_{C^i \in \mathcal{D}^-} \log p(s^i = 0 | d_{R_i-1}, d_{R_i}) \end{aligned} \quad (7)$$

where s^i is 1 for $C^i \in \mathcal{D}$ and s^i is 0 for $C^i \in \mathcal{D}^-$. At the test time, each conversation has a context followed by a set of n candidates responses. The system is tested in its ability to assign a higher score to the true response.

3 Multi-task Crossway Model

Dialogue acts have been shown to be useful for response selection task (Kumar et al., 2018). These dialogue acts can either be given to us or can be predicted using an external model. When dialogue acts are given to us, we denote this model as Siamese-ADA+Crossway. The assumption that the dialogue acts would be available at the test time is rather impractical, therefore an alternate way of leveraging dialogue acts is to predict them. We call this model as Siamese-PDA single-task (Siamese-PDA-ST+Crossway), since dialogue act prediction task is trained independent of the response selection task. In this work we hypothesize that joint modeling of dialogue act prediction task and response selection task would be more beneficial than modeling them individually. Under this hypothesis, we propose a multi-task model, Siamese-PDA-MT+Crossway, that uses the same shared context-response encoder for both tasks. For the dialogue act prediction task of the context and the response, the representations obtained from the shared encoder are input to the classification layer (Section 2.2). The loss is computed as the negative likelihood of predicting the correct dialogue acts for each utterance in the context and the response. The dialogue act prediction loss associated with the context and response are given in Equations (3) and (4) respectively. The response selection task in the multi-task learning setting also uses the same representations as used by the dialogue act prediction task, i.e. those obtained from the shared encoder (Section 2.3). The loss of the response selection task is given in Equation (7). The final loss of the end-to-end multitask model is the combined loss of both the tasks, i.e.

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_r + \mathcal{L}_s$$

4 Experiments

In this section, we provide details of the experiments, i.e. dataset and its preparation, baseline models, experimental setup, results and their analysis including ablation study.

4.1 Datasets

While our model does not need the actual dialogue acts at test time, it does require them at the time of training. So in our problem setting, we require a dataset that is of reasonable size and has utterances annotated with the corresponding dialogue acts. We considered several available datasets (Serban et al., 2015), such as DailyDialog (Li et al., 2017), SwDA (Switchboard Dialogue Act Corpus (Jurafsky, 1997)), MRDA (Meeting Recorder Dialogue Act corpus (Janin et al., 2003)), Ubuntu(Lowe et al., 2015), OpenSubtitles(Tiedemann, 2009), etc. Out of these, Ubuntu, OpenSubtitles, MRDA were found to be unsuitable for our problem setting. The first two, i.e. Ubuntu and OpenSubtitles, do not have dialogue act annotations, and the MRDA corpus is too small, it has only 51 conversations. Therefore, we evaluate the performance of our model on SwDA and DailyDialog datasets:

- SwDA: Switchboard Dialogue Act Corpus (Stolcke et al., 2000) is annotated on 1155 human to human telephonic conversations. Each utterance in a conversation is labeled with one of the 42-class compact DAMSL taxonomy (Core and Allen, 1997; Jurafsky, 1997). The dataset has train, validation, and test splits of 1003, 12, and 19 conversations, respectively.

- DailyDialog(Li et al., 2017) consists of utterances annotated with dialogue acts and is large enough for conversation modeling methods to work. Each utterance is annotated with one of the four dialogue acts. The dataset has train, validation, and test splits of 11118, 1000, and 1000 conversations, respectively.

Dataset Preparation To prepare the data for training and testing, we followed the procedure mentioned in (Kumar et al., 2018; Lowe et al., 2017, 2015). Examples in our training dataset consists of a context of K utterances, followed by the $K+1$ utterance that acts as a true response (positive examples). The training of response selection models require positive examples and an equal number of negative examples. In-order to prepare negative examples, each conversation is replicated by replacing the last utterance with a random utterance from the rest of the training data. On the other hand, examples in our test dataset consists of a context followed by a set of candidate responses of size n. The first response is a true response followed by $n-1$ utterances selected randomly from the rest of the test set. For the DailyDialog and SwDA dataset, the context length K is set to 3 and 10, respectively. For both datasets, the candidate pool consists of 10 utterances. Thus, with

	DailyDialog				SwDA			
	MRR	R@1	R@3	Acc (%)	MRR	R@1	R@3	Acc (%)
Siamese (Lowe et al., 2017)	0.867	0.818	0.880	NA	0.615	0.463	0.687	NA
Siamese-PDA-ST + Crossway	0.900	0.870	0.899	82.9	0.669	0.568	0.685	71.4
Siamese-PDA-MT + Crossway	**0.946**	**0.938**	**0.938**	**86.1**	**0.703**	**0.612**	**0.712**	**72.4**
Siamese-ADA + Crossway (upper bound)	0.956	0.944	0.955	100	0.719	0.630	0.738	100

Table 1: Results for DailyDialog and SwDA Datasets, PDA is for Predicted Dialogue Act, ADA is for Actual Dialogue Act, ST is Single-Task Learning, and MT is for Multi-Task Learning

this data preparation exercise, the total number of conversations in train, test, and validation for DailyDialog are 61030, 2849, and 2695, respectively. And, for the SwDA dataset, the total number number of conversations in train, test, and validation split are 178736, 3483, and 2245, respectively.

4.2 Hyper-parameter Tuning

The validation set is used for fine tuning hyper-parameters, and results are reported on the test set. The maximum batch size is 32; for each batch, the utterances are padded to the maximum length in that batch. We use 300-dimensional Glove embeddings (Pennington et al., 2014) to initialize the word vectors – these word vectors are also updated during training. Both Context-Response encoder and DA-Encoder are GRUs with rnn_size of 300, after optimizing between 100 to 500 in step of 100. Dropout of 0.1 (optimized over 0.0 to 0.7 in steps of 0.1) was applied to embeddings obtained from the output of the encoder. Models were trained to minimize cross entropy using Adam optimizer with learning rate of 0.0003 (optimized over 0.0001, 0.0003, 0.0005, 0.0007, 0.001). All models were trained for 200 epochs.

4.3 Evaluation Metrics

Since our problem formulation is retrieval based, we use standard IR metrics such as Mean Reciprocal Rank (MRR) and Recall@k as our evaluation metrics for the response selection task (Task-2). MRR is calculated as the mean of the reciprocal rank of the true candidate response among other candidate responses. Recall@k measures whether the true candidate response appears in a ranked list of k responses. While we report all of these metrics, in order to make the analysis more explainable, we will keep the MRR as our primary metric. We also report the accuracy of the dialogue act prediction task (Task-1).

4.4 Baseline Methods and Proposed Models

Following is the list of baseline model and proposed models that we use in our experiments:

- **Siamese** (Lowe et al., 2017): a siamese model that uses a dual encoder for conversation modeling without any dialogue acts information.
- **Siamese-PDA-ST+Crossway**: model that uses dialogue acts in single-task setting, (i.e. predicted externally) in a crossway fashion. PDA is for Predicted Dialogue Act and ST is for Single-Task.
- **Siamese-ADA+Crossway**: a hypothetical model that uses actual dialogue acts in a crossway fashion (upper bound). ADA is for Actual Dialogue Act
- **Siamese-PDA-MT+Crossway**: the proposed model uses predicted dialogue acts in a crossway fashion in a multi-task setting. MT is for Multi-Task.
- **Siamese-PDA-MT+Context-DA** (Zhao et al., 2017) : this model uses predicted dialogue acts of the context in a multi-task setting, we implemented this model for the discriminative response selection task.

4.5 Results and Discussion

In Tables 1 and 2, we report results of our experimental study, providing evidences to support two hypotheses:

1. The joint modeling of dialogue act prediction task and response selection task (multi-task setting) performs better than modelling them independently (single-task setting).
2. Combining the dialogue acts of response and context (Crossway) performs better than using either one of them.

	DailyDialogue			SwDA		
	Context-DA (Zhao et al., 2017)	Response-DA	Crossway	Context-DA (Zhao et al., 2017)	Response-DA	Crossway
Siamese-PDA-ST	**0.912**	0.900	0.900	0.639	0.649	**0.669**
Siamese-PDA-MT (Zhao et al., 2017)	0.921	0.919	**0.946**	0.698	0.685	**0.703**
Siamese-ADA (Kumar et al., 2018)	0.934	0.927	**0.956**	0.656	0.682	**0.719**

Table 2: Comparison of MRR when using dialogue acts of Context-only, Response-only and Crossway fashion

Multi-Task vs Single-Task Modelling: In Table 1, we report and compare the results of our proposed method with the baselines, for both datasets, i.e. DailyDialog and SwDA, and provide evidence for the first hypothesis outlined above. From these results, we draw several observations. First observation is that all models that use dialogue acts outperform the model that does not use them. The second observation is that the multi-task model (Siamese-PDA-MT+Crossway), that does the joint modeling of both tasks (dialogue act prediction and response selection task), performs better than the single-task model (Siamese-PDA-ST+Crossway) that models them separately. Multi-task modelling not only improves the MRR in the response selection task for both datasets, but also achieves better dialogue act prediction accuracy. Siamese-ADA+Crossway model which uses the actual dialogue acts, is an upper bound (therefore an ideal model) on how good any model can perform if it were to use predicted dialogue acts. And as we can see, the MRR of multi-task model (Siamese-PDA-MT+Crossway) is close to the upper bounds for both datasets as compared to the single-task model (Siamese-PDA-ST+Crossway). An interesting observation is that, for both DailyDialog and SwDA dataset, though the multi-task model has less than ideal dialogue act prediction accuracy (less than 100%), it performs at par with the ideal model for the response selection task. For the DailyDialog dataset, the multi-task model has the dialogue act prediction accuracy of 86.1%, much less than the ideal accuracy of 100%; in-spite of that, it performs at par with the ideal model that uses the actual dialogue acts, i.e. Siamese-ADA+Crossway (MRR of 0.946 with Siamese-PDA-MT+Crossway vs 0.956 with Siamese-ADA+Crossway). Similarly, for the SwDA dataset, the MRR with multi-task model (Siamese-PDA-MT+Crossway) is 0.703, which is very close to the MRR of 0.719 obtained with the ideal model (Siamese-ADA+Crossway). Consistency of such results across both datasets suggests that the Crossway model is robust and is able to compensate for the errors in predictions by leveraging the similarities across dialogue acts and context/response. In the follow up section, we analyze the effect of Crossway in much more detail.

Crossway vs Response-DA/Context-DA: Although the dialogue acts have been shown to be useful for the response selection task, existing work has only used the dialogue acts of the context. Whereas, in our experiments, we have found that the model that uses the dialogue acts of both context and response outperforms the models that use the dialogue acts of either context or response. To further analyze the results, we perform an ablation study and show the results of using the dialogue acts of context, response and of both. In Table 2, we report the MRR numbers of several models that use the dialogue acts in different settings. More specifically, we show how the following models i.e., Siamese with actual dialogue act (Siamese-ADA), Siamese with predicted dialogue acts in single-task setting (Siamese-PDA-ST) and Siamese with predicted dialogue act in multi-task setting (Siamese-PDA-MT) perform when they are given the dialogue acts of only context (Context-DA), dialogue acts of only response (Response-DA), and dialogue acts of both (Crossway). Results in Table 2 indicate that the Crossway always outperforms the Context-DA or the Response-DA, for both datasets. For DailyDialog dataset, Context-DA performs better than Response-DA for all three models, whereas in the SwDA dataset, Response-DA does a relatively better job than context-DA (two out of three models). Despite their different behavior for different datasets, when we combine Response-DA and Context-DA in a Crossway fashion, it outperforms the both, giving the best of both worlds. This performance improvement of the Crossway over context-DA and response-DA can also be attributed to the

way Crossway model works. Note that in the

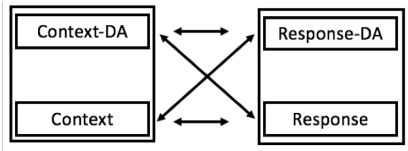

Figure 2: Four implicit similarities being used in Crossway model

Crossway model, there are four similarities that play a role, i.e. context-response, ContextDA-ResponseDA, ContextDA-response and Context-responseDA, graphically depicted in Figure 2. So, in the case of erroneous prediction of either of context DA or response DA, it shall only corrupt two of the four similarities, still leaving two other similarities that can provide strong clues to the underlying model about the correct response belonging to the context.

5 Related Work

Researchers have shown that response selection is a promising approach to build a practical conversation system (Gandhe and Traum, 2010; Lowe et al., 2017; Wu et al., 2016). (Gandhe and Traum, 2010) have shown that response selection based approach for conversation modelling is a good approximation of mimicking human dialogue. Response selection based conversation systems are more practical from the implementation perspective because the responses are mined form previous conversation logs and are therefore more natural and semantically correct. (Ji et al., 2014) have used response selection based techniques for modelling short text conversation responses. They conclude that speech act, sentiment or entity associated with the utterances may enhance the accuracy of the underlying model.

Recently, multi-turn response selection has become the focus of conversation modelling. In multi-turn response selection, current utterance including previous k utterances are used to select an appropriate response from a set of candidate responses. (Lowe et al., 2017; Wu et al., 2016) have shown the efficacy of multi-turn response selection in conversation modeling. (Chaudhuri et al., 2018) have further enhanced these models by incorporating additional domain knowledge in the form of domain specific keywords. (Song et al., 2018) have used an ensemble approach (generation-based and selection-based) to build conversational model. Although effective, none of these methods leverage Dialogue Acts for response selection.

The use of Dialogue Acts (DA) (Xu et al., 2018), latent topics(Zhao et al., 2018; Wen et al., 2017), sentiments, entity models can help in grounding or interpretation of the user utterances which can further aid in improvement of conversation modelling. (Kumar et al., 2018) have shown the usefulness of dialogue act for conversation modeling. However, they assume that the dialogue acts are available at the conversation time which is impractical as dialogue acts are rarely available in a real conversation. Our work addresses this limitation and builds an end-to-end dialogue model, where we predict the dialogue acts and use them as an additional signal for response selection. (Xu et al., 2018; Zhao et al., 2017) use dialogue act for conversation modeling, however the focus of their work is on response generation. In addition to the difference in the underlying task, (Xu et al., 2018) uses an in-house dataset with synthetic dialogue acts [1], whereas we have experimented on publicly available datasets. In (Zhao et al., 2017), while authors propose the model for the dialogue generation task, an important difference is that their model uses the dialogue acts of the context whereas our model uses the dialogue acts of both context and response, combined in a cross-way fashion. We however do use this as a baseline, and show our model's superior performance.

6 Conclusion

This paper presents an end-to-end multi-task model that eliminates the need of actual dialogue acts at the test time. Our end-to-end model combines the predicted dialogue acts of the context and the response with the context, and use the combined representation to select an appropriate response from a set of candidate responses. Our model has been validated on real-world dialogue datasets; we show that our novel way of combining dialogue acts in a cross-way fashion not only compensates for the errors in the dialogue act prediction model but it performs at par with the response selection model that uses actual dialogue acts.

[1]The dataset and the code is not publicly available to be used as a baseline.

References

Debanjan Chaudhuri, Agustinus Kristiadi, Jens Lehmann, and Asja Fischer. 2018. Improving response selection in multi-turn dialogue systems. *arXiv preprint arXiv:1809.03194*.

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.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI Fall Symposium On Communicative Action In Humans And Machines*.

Sudeep Gandhe and David Traum. 2010. I've said it before, and i'll say it again: an empirical investigation of the upper bound of the selection approach to dialogue. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 245–248. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *ICASSP*.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *www. dcs. shef. ac. uk/nlp/amities/files/bib/ics-tr-97-02. pdf*.

Hirotani Kitano. 1991. Phi dm-dialog: an experimental speech-to-speech dialog translation system. *Computer*, 24(6):36–50.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2017. Dialogue act sequence labeling using hierarchical encoder with crf. *arXiv preprint arXiv:1709.04250*.

Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. 2018. Dialogue-act-driven conversation model: An experimental study. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1246–1256.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Ryan Lowe, Nissan Pow, Iulian V Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 285.

Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

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.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Yiping Song, Rui Yan, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, and Dongyan Zhao. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3732–3741. JMLR. org.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Can Xu, Wei Wu, and Yu Wu. 2018. Towards explainable and controllable open domain dialogue generation with dialogue acts. *arXiv preprint arXiv:1807.07255*.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. *arXiv preprint arXiv:1804.08069*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.