Dialogue-act-driven Conversation Model: An Experimental Study

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

The utility of additional semantic information for the task of next utterance selection in an automated dialogue system is the focus of study in this paper. In particular, we show that additional information available in the form of dialogue acts –when used along with context given in the form of dialogue history– improves the performance irrespective of the underlying model being generative or discriminative. In order to show the model agnostic behavior of dialogue acts, we experiment with several well-known models such as sequence-to-sequence encoder-decoder model, hierarchical encoder-decoder model, and Siamese-based models with and without hierarchy; and show that in all models, incorporating dialogue acts improves the performance by a significant margin. We, furthermore, propose a novel way of encoding dialogue act information, and use it along with hierarchical encoder to build a model that can use the *sequential* dialogue act information in a natural way. Our proposed model achieves an MRR of about 84.8% for the task of next utterance selection on a newly introduced DailyDialog dataset, and outperform the baseline models. We also provide a detailed analysis of results including key insights that explain the improvement in MRR because of dialogue act information.

1 Introduction

In the last decade, natural language processing and machine learning –in particular deep learning– have come a long way towards building an automated dialogue system. In a fully automated dialogue system, the goal is to predict an appropriate response given the dialogue history. This problem of response prediction can be formulated in two ways. One is purely generative, where the task is to *generate* a text response, i.e. generating a sentence or utterance from scratch, whereas the other is Next Utterance Selection, where the task is to *select* an appropriate response from a set of given candidates. Despite significant research in text generation, a pure generative model capable of generating syntactically and semantically correct text still remains a distant reality. There have been several efforts such as (Vinyals and Le, 2015; Serban et al., 2016a; Serban et al., 2016b; Serban et al., 2017b) for the task of dialogue generation, however these models still do not seem to work in practice (Liu et al., 2016). This is particularly true for open domain dialogue systems. Dialogue generation in a task-oriented oriented dialogue system, such as flight-booking and troubleshooting, is much easier than in a non-task oriented dialogue system. This level of difficulty arises because a non-task-oriented dialogue system has no predefined goal (or domain), and the vocabulary and possibilities of the dialogues could be endless. Given these challenges, researchers have defined a simpler problem for conversation modeling based on retrieval, i.e. next utterance selection. In this paper we use this second formulation of the problem, and show that using additional information available in the form of dialogue acts help in improving the performance of the underlying model.

Dialogue acts (DA) are higher level semantic abstractions assigned to utterances in a conversation. An example of a dialogue act for an utterance *i'll give you a call tonight* is *Inform* since speaker is providing information. In a traditional dialogue system, where dialogues are formulated by first sentence planning

Proceedings of the 27th International Conference on Computational Linguistics, pages 1246–1256
Santa Fe, New Mexico, USA, August 20-26, 2018.

and then by surface realization, the first step is to understand the dialogue act of the utterance that needs to be generated, and then plan and realize the dialogue accordingly. To better understand the importance of dialogue acts, consider an example of a simple conversation, where if the previous utterance is of type *Question* then the next utterance is most likely going to be of the type, i.e. *Inform*, providing information to that question. Knowing that the next utterance is of type *Inform*, a conversation system with support of dialogue act information can filter a set of candidate responses, and select the most appropriate one. Driven by this intuition, we hypothesize that understanding dialogue acts and using them in the task of next utterance selection should improve the performance irrespective of the underlying model.

Driven by this intuition, we hypothesize that understanding dialogue acts and using them in the task of next utterance selection should improve the performance irrespective of the underlying model.

Most of the existing literature for the task of next utterance selection can be classified into two categories. First is based on Sequence-to-sequence models (generative models) (Serban et al., 2016a; Serban et al., 2017a; Vinyals and Le, 2015), where a model is trained to *generate* a response given context; and the other is Siamese models (discriminative models) (Lowe et al., 2017), where a model is trained to *discriminate* between positive and negative responses for a similar context. In both types of models, at test time, a set of candidate responses is provided consisting of one correct response and several incorrect responses, and the model is evaluated on its ability to assign a higher rank to the true response.

In this paper, through the experimentation with both generative and discriminative types of models, we validate the hypothesis that additional information available in the form of dialogue act significantly improves the performance irrespective of the underlying model. In addition to showing the utility of dialogue acts, we propose a novel model that can use the sequential dialogue act information in a natural way. More specifically, we propose a dialogue-act-driven hierarchical Siamese model. Hierarchical models have shown to perform better than non-hierarchical models for the task of dialogue generation, whereas Siamese models have been shown to outperform the encoder-decoder based models for the task of next utterance selection. In this paper, we combine both of these models, and further enhance them with a dialogue act encoder. The proposed model has a hierarchical encoder which encodes the past utterances, and combine them with the representation of additional contextual information, obtained from the dialogue acts associated with the past utterances, to discriminate the correct response from the incorrect ones. Our proposed model provides us the best of both worlds and outperforms the baseline models by a significant margin. Among others, a key contribution of this paper is that we do a deeper analysis of the reasons for the performance improvement due to inclusion of dialogue act and draw several important key insights such as, dialogue acts induce uniformity in the data, they aid in learning the right patterns. We believe that these insights would inspire new research in this field and push the boundary even further. The main contributions of this paper are as follows:

1. For the task of next utterance selection, we validate the hypothesis that additional information available in the form of dialogue acts improves the performance irrespective of the underlying models.

2. We propose a novel model that combines the strength of Siamese network with strengths of hierarchical structure inherent in the conversations and dialogue act information. The model gives us the best of all, and outperforms the baseline models by a significant margin on the DailyDialog Dataset.

3. We perform a deeper analysis of the utility of the dialogue act information and draw three key insights: models learn dominant dialogue act patterns; dialogue acts induce uniformity; dialogue acts reinforce correct dialogue act patterns.

4. We modify the DailyDialog (Li et al., 2017b) dataset for the task of next utterance selection, and release it publicly along with the code-base of the proposed model[1]. We believe that this dataset will work as a benchmark dataset for further research on this problem. Similar benchmark datasets have been released earlier(Lowe et al., 2015; Serban et al., 2015), however they do not come with dialogue act information.

[1] https://github.com/hk-bmi/ddialog-da-generation

2 Approach

In this section, we provide details of several existing models that we will use to validate our hypothesis. These models include generative models (such as encoder-decoder model and its hierarchical version i.e., hierarchical encoder-decoder) and discriminative model (Siamese-based model). Next, we provide details of the proposed model that adds the hierarchical structure to the Siamese model along with the dialogue act information. To set the notations, we are given a set $\mathcal{D}$ of N conversations, i.e. $\mathcal{D} = (C^1, C^2, \ldots C^N)$, with each conversation C^i being a sequence of R_i utterances, $C^i = (u_1, u_2, \ldots u_{R_i})$. Each utterance u_j in turn is itself a sequence of S_j words, i.e. $u_j = (w_1, w_2, \ldots w_{S_j})$.

2.1 Generative Model

Generative models are the most widely used models for conversation modeling. These models include encoder-decoder model and hierarchical encoder-decoder model.

2.1.1 Encoder-decoder Model

An encoder-decoder is a *generative* model that works on the idea of obtaining a representation of an input and use it for generating an output. It has two main components, encoder and decoder. The encoder encodes the first K utterances, and the decoder uses that encoding to generate the next $K + 1^{th}$ utterance. In a conversation, all words in first K utterances can be stringed together to form a single long chain and passed to an RNN encoder as following:

$$\begin{aligned} e_k &= f^1_{embed}(w_k) \quad \forall k \in 1, 2, \ldots \\ h^e_k &= f^1_{rnn}(h^e_{k-1}, e_k) \quad \forall k \in 1, 2, \ldots \end{aligned} \tag{1}$$

where, f^1_{embed} represents the embedding layer, whereas f^1_{rnn} is the encoder (RNN). Let v be the final output of the encoder which is considered as a representation of the entire context, and used to initialize the decoder (another RNN). Mathematically, the sequence of operations at the decoder are as follows:

$$\begin{aligned} h^d_0 &= v \\ h^d_k &= f^2_{rnn}(h^d_{k-1}, f^2_{embed}(w_k)) \quad \forall k \in 1, 2, \ldots n-1 \\ P_k &= Logistic(h^d_k). \end{aligned} \tag{2}$$

Here, f^2_{embed} represents the embedding layer. $Logistic$ is the final layer, which outputs the probability distribution over the vocabulary. Encoder-decoder models are trained to maximize the likelihood of generating the next utterance, however, for the task of next utterance selection, they are tested based on the probability of generating the candidate utterances.

2.1.2 Hierarchical Encoder-decoder Model

A simple encoder-decoder treats the first K utterances as a single long chain of words, and therefore fails to leverage the hierarchical structure, which is an inherent part of a conversation. Hierarchy is important for conversation modeling since it captures the natural dependency among utterances. Several researchers (Sordoni et al., 2015; Serban et al., 2016b; Serban et al., 2017b; Dehghani et al., 2017; Kumar et al., 2017) have shown that hierarchical models outperform standard non-hierarchical models. Hierarchical models use two encoders to capture the hierarchical structure. The first encoder, referred as utterance encoder, operates at the utterance level, encoding each word in each utterance. The second encoder, referred as conversation encoder, operates at the conversation level, encoding each utterance in the conversation, based on the representations of the previous encoder. These two encoders make sure that the output of the conversation encoder captures the dependencies among utterances. For a given conversation, each word w_k of each utterance u_j is processed by an embedding layer, followed by an RNN which serves as the utterance encoder. Similar to the encoder in equation (1), an utterance encoder gives us a sequence of representations $v_1, v_2, \ldots v_K$, corresponding to the first K utterances $u_1, u_2, \ldots u_K$ in a conversation. These representations are passed on to the conversation encoder, another RNN, which transforms v_j to another representation g_j. The representation obtained from the last time-step of the

conversation-level encoder i.e. g_K is considered as the representation of the entire conversation and used to initialize the decoder which works in the same way as Equation 2.

2.2 Discriminative Model

A decoder in the encoder-decoder model generates the next word given the context, and though it has several valid and reasonable choices, it is burdened with the task of generating exactly a particular choice that matches the ground truth. For example, for a context *I am enjoying the day, it is warm and sunny*, if decoder generates *yes, it is.* and the ground truth dictates *yes, indeed, it is a lovely day*, the decoder has failed, though it is a valid response. Due to these challenges with generative models, discriminative models are trained directly to discriminate between positive and negative utterances. A typical discriminative model, or in particular Siamese model, consists of two encoders, one encoder encoding the context, while another encoding the candidate utterance, i.e utterance $K+1$. These two representations are passed to a final layer that computes the probability of candidate being a valid response given the context. Let $h^{(1)}$ and $h^{(2)}$ be the representations obtained from the first encoder and second encoder, respectively, then the probability of their association can be computed using the following expression.

$$p(s|h^{(1)}, h^{(2)}) = \sigma\big({h^{(1)}}^T A h^{(2)} + b\big) \tag{3}$$

where, the bias b and matrix A are learned model parameters.

2.3 Dialog-act-driven Models

Dialogue acts are higher level abstractions assigned to utterances. In our problem setting, we are given a list of dialogue acts $da_1, da_2, \ldots da_K$, corresponding to first K utterances in the conversation. These dialogue acts are treated as an additional sequence of signals that can aid in the learning process, and are passed through an encoder, denoted as Dialog-Act encoder (DA-encoder). The DA-encoder works on the same principle as the utterance encoder. It builds a dialogue act vocabulary and uses that to learn dialogue act embeddings. Similar to the utterance encoder, the input to the DA-encoder are one hot encodings of the dialogue acts, which are then passed through an embedding layer to learn DA embeddings. These DA embeddings are sent to an RNN to learn dialogue act representations. The sequence of operations for the DA-encoder are as follows:

$$\begin{aligned} e_{da_k} &= f^3_{embed}(da_k) \quad \forall k \in 1, 2, \ldots K \\ h_{da_k} &= f^4_{rnn}(h_{da_{k-1}}, e_{da_k}) \quad \forall k \in 1, 2, \ldots K. \\ q_K &= h_{da_K} \end{aligned} \tag{4}$$

The output of the DA-encoder at the last time step (q_K) gives us the representation of the entire DA sequence which is then used in the further modeling process in generative and discriminative models. In generative models, it is used in the decoder by concatenating g_K and q_K, whereas in discriminative models, it is used along with encoder's output by combining g_K with q_K through a linear combination.

2.4 Dialog-act-driven Hierarchical Siamese - Proposed Model

Our proposed model, i.e. Dialog-act-driven Hierarchical Siamese Model (HSiamese-DA), uses the following three components: a hierarchical encoder to obtain a representation that captures the dependencies among K utterances; an utterance encoder to obtain a representation of the candidate response, $(K+1)^{th}$ utterance; a DA-encoder (Equation 4) that captures the dependencies among the dialogue acts of the first K utterances. Let the representation obtained from the hierarchical encoder, DA-encoder and utterance encoder be g_K, q_K and v_{K+1}, respectively. The two representations, g_K and q_K, are linearly combined to obtained a compositional representation of the context, which is then used along with candidate representation to compute the probability of associating the candidate response with the context using following expression:

$$\begin{aligned} d_K &= \alpha * g_K + (1-\alpha) * q_K \\ p(s|d_K, v_{K+1}) &= \sigma(d_K^T \cdot A \cdot v_{K+1} + b) \end{aligned} \tag{5}$$

The model is trained by minimizing the cross-entropy of all labeled conversations including positive and negative examples. At the test time, each conversation has K utterances followed by a set of 10 candidates responses. The system is tested in its ability to assign a higher rank to the true response.

3 Experiments

In this section, we describe the details of the experiments, i.e. dataset and its preparation, baseline models, experimental setup, and analysis of results.

3.1 Dataset

In our problem setting, we require a dataset that is of reasonable size[2] and has utterances annotated with the corresponding dialogue acts. Although there are several available datasets(Serban et al., 2015), such as 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., they are not really suitable for our problem setting. Most of these datasets do not come with dialogue acts, and the ones which do (i.e. SWDA and MRDA) are small in size. Note that the SwDA and MRDA datasets contain 1003 and 51 conversations, respectively. To the best of our knowledge, a recently released dataset, DailyDialog(Li et al., 2017b), is the only dataset that has utterances annotated with dialogue acts and is large enough for conversation modeling methods to work. Furthermore, in this dataset, conversations are non-task oriented, and each conversation focuses on one topic. Each utterance is annotated with four dialogue acts as described in Table 1. The dataset has train, validation, and test splits of 11118, 1000, and 1000 conversations, respectively. We evaluate and report our results on the DailyDialog dataset.

In this paper, we hypothesize that dialogue acts improve conversation modeling. However, it is not always possible that such dialogue acts are available in practice, and it would be ideal to predict dialogue acts first (Kumar et al., 2017), and then use them for next utterance generation/retrieval; having a model where both tasks, i.e. prediction and generation, are performed simultaneously may not be ideal for validating the hypothesis. Note that the error from the dialogue act prediction may propagate to the next utterance generation/retrieval. Therefore, we intentionally did not use the predicted dialogue acts (rather used the available dialogue acts) to make sure that the insights about the usefulness of the dialogue acts are not corrupted due to the error in the upstream prediction model.

Dialogue Act	Description
Inform	A speaker is providing information by means of a question or statement
Question	A speaker intends to obtain information by asking a question
Directive	A speaker is requesting, accept/reject offer, or making a suggestion
Comissive	A speaker accept/reject a request or suggestion

Table 1: Dialogue Acts and their description available in the DailyDialog Dataset.

3.1.1 Dataset Preparation for Next Utterance Selection Task

The DailyDialog dataset in its original form is not directly useful for the task of next utterance selection, and hence requires preparation. The dataset has the dialogues from both the speakers. Owing to the different conversational style of human and conversation agent, our objective is to build a model that is specific to the agent, i.e. bot. Therefore, we need to modify the dataset in such a way that we only consider those turns where we need to predict the bot's utterance. To clarify further, consider the example conversation given in Table 2. The conversation has 8 utterances, and each utterances is marked with the speaker, i.e. human (H) and bot (B). Since we are only interested in building bot-specific model, we only pick those subsequences from this conversation where the last utterance is "B". This gives us three subsequences: 1,2,3,4; 3,4,5,6; 5,6,7,8 for a context of size 3. In each of these sub-conversations,

[2]Generative models such as sequence-to-sequence or discriminative models such as Siamese only perform better when there is large amount of data for training.

the first three utterances constitute the context, while the last utterance is the true response. Our training data consists of such subsequences made up of 4 utterances. In the test data, each subsequence, in addition to these 4 utterances, has 9 more utterances selected randomly from the test pool, therefore a total of 13 utterances. These 9 utterances along with the 4^{th} response (i.e., true response) utterance constitute the candidate pool. With this data preparation exercise, the total number of conversations in train, test, and valid are 30515, 2849, and 2695, respectively. This version of dataset is used for training and testing generative models. For discriminative models, data required is a bit different. The training of discriminative models require positive examples and an equal number of negative examples. Note that training data of the generative models did not have any negative examples. In-order to prepare negative examples, we replicate each conversation by replacing the last utterance with a random utterance from the test data. The test and valid dataset remain as in the generative models. Thus, with this data preparation exercise, the total number of conversations in train, test, and valid are 61030, 2849, and 2695, respectively.

Id		Utterance	DA
1	H	Hello, this is Mike, Kara .	I
2	B	Mike! Good to hear from you. How are you?	Q
3	H	Everything is fine, and how are you?	Q
4	B	Things are going well with me.	I
5	H	Kara, I had fun the other night at the movies and was wondering if you would like to go out again this Friday.	D
6	B	Mike, I don't think that it's a good idea to go out again.	C
7	H	Maybe we could just meet for coffee or something.	D
8	B	I can't really deal with any distractions right now, but I appreciate the nice evening we spent together.	C

Table 2: A snippet of a conversation showing few dialogues between a Human (H) and Bot (B).

3.2 Baseline models and Proposed Model

Here we list the baseline models, their modified version enhanced with dialogue act information, and the proposed model.

Generative Models:

- **ED** - It is a vanilla sequence to sequence model that uses an utterance encoder to obtain a representation of first K utterances which is then used in a decoder to generate next utterance.
- **HRED** - An extension of sequence to sequence model that uses a hierarchical encoder to obtain a representation of first K utterances, which is then used in decoder to generate next utterance.
- **ED-DA** - An extension of the ED model which uses dialogue act information. It has a conditional decoder, that conditions the generation of each word on the dialogue acts representation.
- **HRED-DA** - An extension of the HRED model which uses dialogue act information. Similar to ED-DA, it also has a conditional decoder that conditions the generation of each word on the dialogue acts representation.

Discriminative Models

- **Siamese** - Also known as Dual-Encoder, it uses two encoders (both utterance encoders) with shared weights, to produce the representation for the K utterances and the $(K+1)$ utterance.
- **HSiamese** - A Hierarchical version of the Siamese model that uses a hierarchical encoder to produce a representation for the K utterances, and a plain encoder (utterance encoder) to produce a representation for the $(K+1)$ utterance.

- **Siamese-DA** - An extension of Siamese model that uses the additional dialogue act information obtained through DA-encoder. The representation obtained from the DA-encoder is linearly combined with the representation of the K utterances obtained from an utterance encoder.
- **HSiamese-DA** - The proposed model uses a Hierarchical Encoder and a DA-Encoder. The representation obtained from the DA-Encoder is linearly combined with the representation obtained from the hierarchical encoder.

3.3 Hyper-parameter Tuning

In our experiments, the parameters are tuned on validation set while the results are reported on test set. Each utterance in a mini-batch was padded to the maximum length for that batch. The maximum batch size allowed was 32. The word vectors were initialized with the 300-dimensional Glove embeddings (Pennington et al., 2014), and were also updated during training. For the generative models, the utterance encoder, conversation encoder, DA-Encoder and decoder are all GRUs with rnn_size set to 1000 (optimized over 100 to 1200 in steps of 100). For the discriminative model, the utterance encoder, conversation encoder, and DA-Encoder are all GRUs with rnn_size set to 300 (optimized over 100 to 500 in steps 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 conversation encoder. Note that, dropout was not used in the discriminative model and its variations. 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). We found that a higher learning rate up-to 0.0005 helps the model to learn quickly, whereas learning rate greater than 0.0005 leads to oscillations.

3.4 Results and Discussion

In this section, we present results of our experimental study, followed by its analysis.

3.4.1 Performance Evaluation

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. 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.

In this work, our hypothesis is that additional information about utterances available in the form of dialogue acts helps irrespective of the underlying model, i.e. generative or discriminative. Results in Table 3 support our hypothesis. These results clearly indicate that the MRR of the true candidate response improves when dialogue acts of previous utterances are provided. From these tables we see that for all underlying models, the dialogue act version performs better than non dialogue act version. These results furthermore indicate that hierarchical version performs better than non-hierarchical version for both generative and discriminative models. In the generative case, the plain ED has an MRR of 0.474, whereas the same model, when conditioned with DA-Encoder, has an MRR of 0.54, an improvement of 13.9%. The hierarchical encoder-decoder HRED and HRED-DA has an MRR of 0.523 and 0.583, respectively, an improvement of 11.4%. Generative models are sequence-to-sequence models and rather complex in nature, so it is interesting to note that even a much simpler discriminative model, i.e. plain Siamese model, without any dialogue act information, has an MRR of 0.8 compared to 0.58 of the best performing generative model, i.e. HRED-DA. This observation demonstrates the strength of the discriminative models, and therefore is a motivation behind the proposed model. The proposed model improves these baseline numbers by incorporating hierarchy and dialogue act information, and pushes the MRR to 0.848.

3.4.2 Performance Analysis

While we have shown that using dialogue act information does help in the next utterance selection task, in this section, we dig deeper and understand reasons for it. In order to do that, we analyze the dialogue act distribution of the test data and model outputs. Although all K dialogue acts corresponding to K utterances in the context might play a role in ranking candidate utterances, the following analysis only

	MRR	R@1	R@2	R@5
ED	0.474	0.327	0.405	0.639
ED-DA	0.54	0.407	0.478	0.690
HRED	0.523	0.384	0.471	0.676
HRED-DA	0.583	0.448	0.542	0.742

(a) Generative Models

	MRR	R@1	R@2	R@5
Siamese	0.800	0.711	0.785	**0.949**
Siamese-DA	0.844	0.784	**0.824**	0.944
HSiamese	0.817	0.743	0.792	0.948
HSiamese-DA	**0.848**	**0.795**	0.821	0.932

(b) Discriminative Models

Table 3: Comparison of different models with and without dialogue acts

uses the pairs of dialogue acts, i.e. dialogue acts of K^{th} and $(K+1)^{th}$ utterances. Tables 4(a), 4(b) and 4(c) show the distribution of such dialogue act pairs for test data, HSiamese, and HSiamese-DA models respectively. Here, rows indicate the dialogue act of K^{th} utterance, whereas columns indicate the dialogue act of $(K+1)^{th}$ utterance. A cell value indicates the count of utterance pairs with the respective dialogue act combinations where $(K+1)^{th}$ utterance was ranked 1. Note that in the test data, $(K+1)^{th}$ utterance is the true candidate response and always have the rank 1. For instance, there are 742 utterance pairs in the test data, where K^{th} and $(K+1)^{th}$ utterances have dialogue acts Q and I, respectively, however, out of those 742 instances, HSiamese ranked only 605 as 1 while HSiamese-DA ranked 638 as 1. From these tables, we draw following observations.

	I	Q	D	C
I	600	336	176	16
Q	742	73	87	2
D	26	75	56	305
C	69	50	76	6

(a) Ground-truth test data

	I	Q	D	C	R@1
I	403	208	115	9	0.65
Q	605	34	64	2	0.78
D	16	41	42	192	0.63
C	48	33	56	4	0.70

(b) HSiamese

	I	Q	D	C	R@1
I	446	253	121	12	0.74
Q	638	43	61	2	0.82
D	17	58	42	243	0.78
C	52	39	68	5	0.82

(c) HSiamese-DA

Table 4: Number of rank-1 conversations and theie DAs for K^{th} and $(K+1)^{th}$ utterances.

Models Learn Dominant Patterns: The first is that there are certain dominant communication patterns that we observe in both, test data and model outputs (See Table 4), suggesting that models are able to learn these patterns and retain them in their outputs . We observe that a *Question* is often followed by an *Information*, whereas an *Information* can be followed by another *Information* or a *Question*. A *Directive* tends to be followed by *Commissive*. These communication patterns not only make sense intuitively but they are also in agreement with previous studies (Li et al., 2017b; Ribeiro et al., 2015).

Dialogue Acts Bring Uniformity: The second and a rather more important observation is that dialogue acts help the most for the dialogue act class (DA-class) when the utterances belonging to that class are non-uniform in their linguistic construct. In order to better exlpain this, we first compute the break-up of recall@1 according to the dialogue act classes. A DA-class of a conversation in the test data is defined based on the dialogue act of the last utterance (K^{th} utterance) in the context. These numbers are shown in the last column of Tables 4(b) and 4(c) for the respective models. In Table 4(b), first row in recall@1 column is 0.65, which indicates that out of the total number of test conversations where dialogue act of the last utterance of context was I, 65% of true candidate responses were ranked 1 by the HSiamese model. Such a DA-class wise breakup of the recall@1 numbers helps us do an analysis with respect to individual DA-classes. From this break-up, it is clear that for the HSiamese model, *Question* DA-class has the best performance of 78% whereas *Directive* has the worst performance of 63%. This difference can be attributed to the fact that all utterances with dialogue act as *Question* have rather uniform construct. Some examples of *Question* utterances are, '*Q: Do you have a fever?*' and '*Q: Why do you want to work for our company?*', while the examples of *Directive* utterances are, '*D:when we have the final results, we will call you.*' and '*D:we will take the trip. could you give us a pamphlet?*'. From these examples, we observe that utterances belonging to DA-class *Question* have rather uniform construct in terms of linguistic features, whereas utterances belonging to DA-class *Directive* are ambiguous –some of the utterances of type *Directive* can be easily confused for *Question*. This uniformity makes the learning task easier for *Question* class, and thereby giving us better results in the next utterance selection

task, even for the model that does not use the dialogue act information. This performance difference reduces when we provide the dialogue act information along with the textual content (See Table 4(c)). Inclusion of dialogue act information within a non-uniform class such as *Directive* brings in uniformity, and therefore, results in significant performance improvement. In the case of *Directive*, we see as much as 15 percentage point improvement. Similar improvement pattern has been observed for DA-class *Information* and *Comissive*. Similar to *Directive*, utterances belonging to *Comissive* have rather non-uniform construct, and with the availability of dialogue acts, this DA-class is able to gain much more than the *Information* DA-class.

	I	Q	D	C	Total
I	43	45	6	3	97
Q	33	9	-3	0	39
D	1	17	0	51	69
C	4	6	12	1	23
Total	81	77	15	55	228

Table 5: Difference of number of rank-1 conversations between HSiamese-DA and HSiamese.

Dialogue Acts Help Model Learn the Right Patterns: In Table 5, we show the relative improvement of HSiamese-DA model over HSiamese. From this table, we observe that there are a total of 228 conversations where the proposed model was able to improve the ranking of true candidate response to 1. We further observe that the biggest improvement is in $I \rightarrow I$, $I \rightarrow Q$, $Q \rightarrow I$, and $D \rightarrow C$, which make sense intuitively. These are dominant patterns observed in the training data which should be preserved in the model output as well, however these patterns will only be preserved when model is able to capture the *correct* dialogue act information. Since in many cases D and Q have similar construct, without explicit dialogue act information, a model may get confused and may learn patterns not observed in the training data. For example, $Q \rightarrow I$ and $D \rightarrow C$ are the dominant and right patterns in the training data, however in the absence of explicit dialogue act information, the model may get confused between D and Q and may learn $D \rightarrow I$ and $Q \rightarrow C$ instead of the dominant patterns i.e. $Q \rightarrow I$ and $D \rightarrow C$. With the explicit dialogue act information, this ambiguity is alleviated and model learns the right patterns as demonstrated by Table 5. Similar observations are true for other two constructs, i.e. *Information* and *Commisive*. Both are rather similar in construct, '*I: No, thank you*', '*I: It doesn't matter. it happens to everyone .*' and '*C: I knew you'd see it my way .*', '*C: Ok , i am ready to think of other things .*', and there is no obvious distinguishing factor. However, providing explicit DA information helps in disambiguation, and learn the patterns that are observed in the training data such as $I \rightarrow I$, $I \rightarrow Q$.

4 Related Work

In conversation modeling, the most basic problem is to generate a response given a context. Several efforts have been made towards solving the problem of dialogue generation(Vinyals and Le, 2015; Liu et al., 2016; Li et al., 2015), however, due to the inherent difficulty of the problem, these efforts have only had limited success and are known to have issues like generating repetitive and generalized responses such as *I don't know* or *Ok*.

For the task of Next Utterance Selection, which is a relatively simpler problem than generation, though existing generative models can be easily adopted, their counterpart discriminative models have shown to have better performance. In generative models, the most notable work is from (Vinyals and Le, 2015), however this work considers the context as a flat long string of words and ignores the hierarchical structure. Researchers have proposed hierarchical model (Serban et al., 2016b) and their variations (Serban et al., 2017b; Serban et al., 2017a; Li et al., 2017a) but none of these models take into account the dialogue act information. In Discriminative models, such as Siamese, a very notable work by (Kannan et al., 2016), *smart reply*, retrieves the most likely response from a set of candidate response clusters. (Lowe et al., 2017) has used a retrieval based Siamese model and shown its results on the Ubuntu corpus. Our

proposed model builds upon the strengths of generative and discriminative models, and uses hierarchy along with the dialogue act information to achieve the best performance. A recent work by (Zhao et al., 2017) has used dialogue acts for the task of dialogue generation. Our work complements their findings, and further show that dialogue acts improve the model performance across the board irrespective of underlying model (i.e. generative or discriminative models) and for the task of next utterance selection.

5 Conclusion

For the task of next utterance selection, we show that dialogue acts helps achieve better performance irrespective of the underlying model, be it generative or discriminative. We also propose a novel discriminative model that leverages the hierarchical structure in a conversation and dialogue act information to produce much improved results, an MRR of 0.848. Our results not only show the improvement in performance, but we also present key reasons for it by doing a detailed analysis and drawing key insights that the inclusion of dialogue act information induces uniformity and removes ambiguity.

References

Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session based query suggestion. In *CIKM*.

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*.

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*.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 955–964. ACM.

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*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. In *NAACL*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. 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*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2015. The influence of context on dialogue act recognition. *arXiv preprint arXiv:1506.00839*.

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*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016a. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C Courville. 2017a. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI*, pages 3288–3294.

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

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *CIKM*.

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.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 654–664.