Dialogue Coherence Assessment Without Explicit Dialogue Act Labels

Mohsen Mesgar **Sebastian Bücker** **Iryna Gurevych**

Ubiquitous Knowledge Processing Lab (UKP)
Technische Universität Darmstadt (TUDa)
{mesgar,buecker,gurevych}@ukp.tu-darmstadt.de

Abstract

Recent dialogue coherence models use the coherence features designed for monologue texts, e.g. nominal entities, to represent utterances and then explicitly augment them with dialogue-relevant features, e.g., dialogue act labels. It indicates two drawbacks, (a) semantics of utterances is limited to entity mentions, and (b) the performance of coherence models strongly relies on the quality of the input dialogue act labels. We address these issues by introducing a novel approach to dialogue coherence assessment. We use dialogue act prediction as an auxiliary task in a multi-task learning scenario to obtain informative utterance representations for coherence assessment. Our approach alleviates the need for explicit dialogue act labels during evaluation. The results of our experiments show that our model substantially (more than 20 accuracy points) outperforms its strong competitors on the DailyDialogue corpus, and performs on par with them on the SwitchBoard corpus for ranking dialogues concerning their coherence. We release our source code[1].

1 Introduction

Considering rapid progresses in developing open-domain dialogue agents (Serban et al., 2016; Ghazvininejad et al., 2018; Dinan et al., 2019; Li et al., 2019), the need for models that compare these agents in various dialogue aspects becomes extremely important (Liu et al., 2016; Dinan et al., 2019). Most available methods for dialogue evaluation rely on word-overlap metrics, e.g. BLEU, and manually collected human feedback. The former does not strongly correlate with human judgments (Liu et al., 2016), and the latter is time-consuming and subjective. A fundamental aspect of dialogue is coherence – what discriminates a high-quality dialogue from a random sequence of dialogue utterances (Halliday and Hasan, 1976; Grosz and Sidner, 1986; Byron and Stent, 1998). Dialogue coherence deals with semantic relations between utterances considering their dialogue acts (Perrault and Allen, 1978; Cervone et al., 2018).

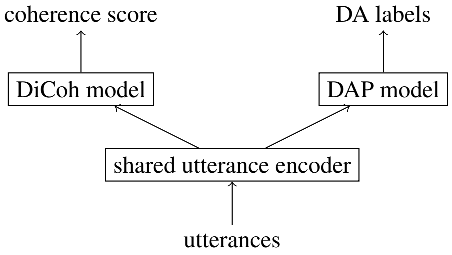


Figure 1: A high-level view of our multi-task learning approach for dialogue coherence modeling.

A Dialogue Act (henceforth *DA*) gives a meaning to an utterance in a dialogue at the level of "illocutionary force", and therefore, constitutes the basic unit of communication (Searle, 1969; Raheja and Tetreault, 2019). A DA captures what a speaker's intention is of saying an utterance without regard to the actual content of the utterance. For example, a DA may indicate whether the intention of stating an utterance is to ask a question or to state a piece of information.

Recent approaches to dialogue coherence modeling use the coherence features designed for monologue texts, e.g. entity transitions (Barzilay and Lapata, 2005), and augment them with dialogue-relevant features, e.g., DA labels (Cervone et al., 2018). These DA labels are provided by human annotators or DA prediction models. Such coherence models suffer from the following drawbacks: (a) they curb semantic representations of utterances to entities, which are sparse in dialogue because of short utterance lengths, and (b) their performance relies on the quality of their input DA labels.

[1] https://github.com/UKPLab/acl2020-dialogue-coherence-assessment

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1439–1450
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

We propose a novel approach to dialogue coherence assessment by utilizing dialogue act prediction as an auxiliary task for training our coherence model in a multi-task learning (MTL) scenario (Figure 1). Our approach consists of three high-level components: *an utterance encoder*, *a dialogue coherence model (DiCoh)*, and *a Dialogue Act Prediction (DAP) model*. The layers of the utterance encoder are shared between the DAP and the DiCoh model. This idea enables our DiCoh model to learn to focus on salient information presented in utterances considering their DAs and to alleviate the need for explicit DA labels during coherence assessment.

We evaluate our MTL-based approach on the DailyDialog (Li et al., 2017) and SwitchBoard (Jurafsky and Shriberg, 1997) English dialogue corpora in several discriminating experiments, where our coherence model, DiCoh, is examined to discriminate a dialogue from its perturbations (see Table 1). We utilize perturbation methods, like *utterance ordering* and *utterance insertion*, inherited from coherence evaluation approaches for monologue texts, and also introduce two dialogue-relevant perturbations, named *utterance replacement* and *even utterance ordering*.

Our core contributions are: (1) proposing an MTL-based approach for dialogue coherence assessment using DAP as an auxiliary task, yielding more informative utterance representations for coherence assessment; (2) alleviating the need for DA labels for dialogue coherence assessment during evaluations; (3) an empirical evaluation on two benchmark dialogue corpora, showing that our model substantially outperforms the state-of-the-art coherence model on DailyDialog, and performs on par with it on SwitchBoard.

2 Related Work

Early approaches to dialogue coherence modeling are built upon available models for monologue, such as the EntityGrid model (Barzilay and Lapata, 2005, 2008). EntityGrid and its extensions (Burstein et al., 2010; Guinaudeau and Strube, 2013; Mesgar and Strube, 2014; Tien Nguyen and Joty, 2017; Farag and Yannakoudakis, 2019) rely on entity transitions, as proxies of semantic connectivity, between utterances. These approaches are agnostic to discourse properties of dialogues (Purandare and Litman, 2008; Gandhe and Traum, 2008; Cervone et al., 2018).

	Utterance	DA label
coherent		
utt_1:	*This is my uncle, Charles.*	inform
$\mathbf{utt_2}$	***He looks strong. What does he do?***	question
utt_3:	*He's a captain.*	inform
$\mathbf{utt_4}$:	***He must be very brave.***	inform
utt_5:	*Exactly!*	inform
incoherent		
utt_1::	*This is my uncle, Charles.*	inform
$\mathbf{utt_4}$:	***He must be very brave.***	inform
utt_3:	*He's a captain.*	inform
$\mathbf{utt_2}$:	***He looks strong. What does he do?***	question
utt_5:	*Exactly!*	inform

Table 1: An example dialogue from DailyDialog (top) and its perturbation (bottom), which is generated by permuting the utterances said by one of the speakers (shown in boldface), and is less coherent. The right column shows the DA labels associated with utterances.

Inspired by EntityGrid, Gandhe and Traum (2016) define transition patterns among DA labels associated with utterances to measure coherence. Cervone et al. (2018) combine the above ideas by augmenting entity grids with utterance DA labels. This model restricts utterance vectors only to entity mentions, and needs gold DA labels as its inputs for training as well as evaluation. However, obtaining DA labels from human annotators is expensive and using DAP models makes the performance of coherence model dependent on the performance of DAP models.

Recent approaches to dialogue coherence modeling benefit from distributional representations of utterances. Zhang et al. (2018) quantify the coherence of dialogue using the semantic similarity between each utterance and its preceding utterances. This similarity is estimated, for example, by the cosine similarity between an utterance vector and a context vector where those vectors are the average of their pre-trained word embeddings. Vakulenko et al. (2018) measure dialogue coherence based on the consistency of new concepts introduced in a dialogue with background knowledge. Similarly, Dziri et al. (2019) utilize a natural language inference model to assess the content consistency among utterances as an indicator for dialogue coherence. However, these approaches lack dialogue-relevant information to measure coherence.

Our MTL-based approach solves these issues: (i) it benefits from DAs and semantics of utterances to measure dialogue coherence by optimizing utterance vectors for both DAP and coherence assessment, and (ii) it uses DA labels to define an auxiliary task for training the DiCoh model using MTL, instead of utilizing them in a pipeline. Therefore, it efficiently mitigates the need for explicit DA labels as inputs during coherence assessment.

3 Method

We represent a dialogue between two speakers as a sequence of utterances, $\mathit{dial} = [\mathit{utt}_1, ..., \mathit{utt}_m]$. We address the problem of designing a coherence model, DiCoh, which assigns a coherence score to *dial*, $s_{\mathit{dial}} = \mathit{DiCoh}(\mathit{dial})$. Given a pair of dialogues $\phi = (\mathit{dial}_i, \mathit{dial}_j)$, our *DiCoh* model ideally assigns $s_{\mathit{dial}_i} > s_{\mathit{dial}_j}$ if and only if dialogue dial_i is preferred over dialogue dial_j according to their perceived coherence. Instead of using gold DA labels as inputs to DiCoh, we use them to define an auxiliary task and model, DAP, to enrich utterance vectors for DiCoh in an MTL scenario. Figure 2 shows a low-level illustration of our MTL-based approach.

Utterance encoder We use a word embedding layer, *Emb*, to transform the words in utterance $\mathit{utt} = [w_1, ..., w_n]$ to a sequence of embedding vectors $E = [e_1, ..., e_n]$, where n is the number of words in *utt*. The embedding layer can be initialized by any pre-trained embeddings to capture lexical relations. We use a Bidirectional recurrent neural network with Long Short-Term Memory cells, *BiLSTM*, to map embeddings E to encode words in their utterance-level context:

$$\begin{aligned} E &= \mathit{Emb}(\mathit{utt}), \\ H_u &= \mathit{BiLSTM}(E), \end{aligned} \tag{1}$$

where H_u shows the hidden state vectors $[h_1^u, ..., h_n^u]$ returned by *BiLSTM*. At word t, h_t^u is the concatenation of hidden states of the forward $\overrightarrow{h_t^u}$ and the backward LSTMs $\overleftarrow{h_t^u}$:

$$h_t^u = [\overrightarrow{h_t^u}; \overleftarrow{h_t^u}]. \tag{2}$$

We apply a self-attention mechanism, *Atten*, to the hidden state vectors in H_u to obtain the vector representation, u, of utterance utt:

$$u = \mathit{Atten}(H_u). \tag{3}$$

Generally, the attention layer, *Atten*, for an input vector x is defined as follows:

$$\begin{aligned} \beta_t &= x_t * W, \\ \alpha_t &= \frac{\exp(\beta_t)}{\sum_t \exp(\beta_t)}, \\ o &= \sum_t \alpha_t * x_t, \end{aligned} \tag{4}$$

where W is the parameter of this layer, and o is its weighted output vector. Attention enables the utterance representation layer to encode an utterance by the weighted sum of its word embeddings. It is worth noting that the parameters of the utterance encoder are shared for representing all utterances in a dialogue.

DiCoh model For an input dialogue $\mathit{dial} = [\mathit{utt}_1, ..., \mathit{utt}_m]$, the output of the utterance representation encoder is a sequence of vectors, i.e., $[u_1, ..., u_m]$. Our coherence assessment model (DiCoh) combines these vectors by a *BiLSTM* to obtain dialogue-level contextualized representations of utterances. Then, a self-attention (Equation 4) with new parameters computes the weighted average of the contextualized utterance vectors to encode the dialogue:

$$\begin{aligned} [h_1^d, ..., h_m^d] &= \mathit{BiLSTM}([u_1, ..., u_m]), \\ d &= \mathit{Atten}([h_1^d, ..., h_m^d]). \end{aligned} \tag{5}$$

A linear feed-forward layer, FF, maps the dialogue vector, d, to a dialogue coherence score, s_{dial}:

$$s_{\mathit{dial}} = \mathit{FF}(d). \tag{6}$$

DAP model Our DAP model, which is used to solve the auxiliary DAP task, is a $softmax$ layer which maps an utterance vector, u, to a probability distribution p_a over DA labels A:

$$p_a(u) = \mathit{softmax}(W_{|u|\times|A|} * u + b), \tag{7}$$

where $W_{|u|\times|A|}$ shows the weights of the *softmax* layer, $|u|$ is the size of the utterance vector, $|A|$ is the number of DA labels, and b is the bias.

3.1 Multi-Task Learning

As illustrated in Figure 1, our main idea is to benefit from the DAP task for improving the performance of the dialogue coherence model by using them in a multi-task learning scenario. We also assume that each utterance utt_k is associated with DA label, a_k, during training but not during evaluation.

We define a loss function for each task, and then use their weighted average as the total loss. The DAP loss function for dialogue $dial$ is the average cross-entropy:

$$\mathcal{L}_{da}^{\mathit{dial}} = -\frac{1}{m} \sum_{k\in(1,...,m)} a_k * \log(p_a(u_k)), \tag{8}$$

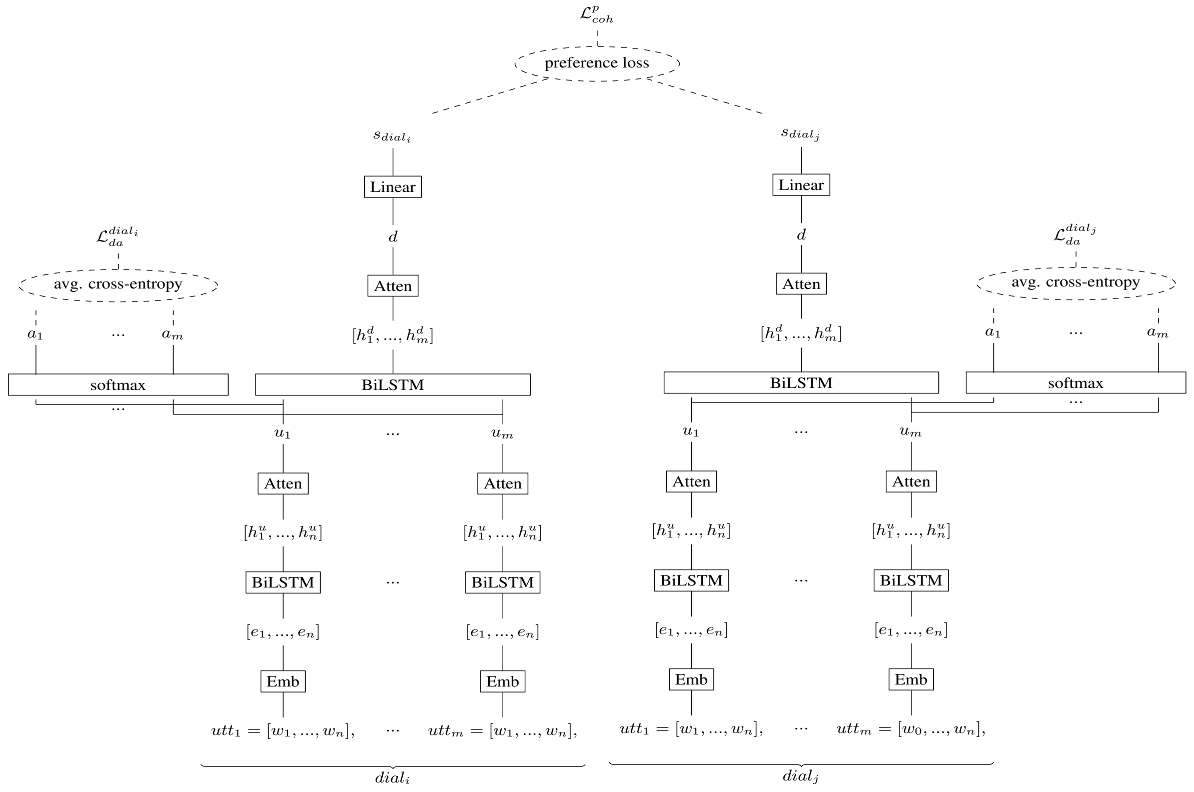


Figure 2: A low-level illustration of our MTL-based approach to dialogue coherence assessment. The input is dialogue pair $p = (dial_i, dial_j)$. Dashed items represent losses. Models' parameters are shared among dialogues.

where m is the number of utterances in dialogue, and a_k is the one-hot vector representation of the gold DA label associated with the k^{th} utterance. $log(p_a)$ is the natural log of probabilities over DA labels, which is obtained in Equation 7.

Inspired by preference learning approaches (e.g. the proposed method by Gao et al. (2019) for text summarization) we define the loss function for coherence assessment through pairwise comparisons among dialogues. Given dialogue pair $\phi = (dial_i, dial_j)$ and its preference coherence label,

$$l^c = \begin{cases} 0 & \text{if } dial_i \text{ is preferred over } dial_j, \\ 1 & \text{otherwise,} \end{cases} \tag{9}$$

the coherence loss is:

$$\mathcal{L}^{\phi}_{coh} = \max\{0, 1 - s_{\phi[l^c]} + s_{\phi[1-l^c]}\}, \tag{10}$$

where $[.]$ is the indexing function. More formally, $s_{\phi[l^c]}$ and $s_{\phi[1-l^c]}$ are the coherence scores of the coherent and incoherent dialogue in pair $\phi = (dial_i, dial_j)$, respectively. Finally, the total loss value is the weighted combination (Kendall et al., 2018) of the above losses:

$$\mathcal{L} = \frac{\mathcal{L}^{\phi}_{coh}}{\gamma_1^2} + \frac{(\mathcal{L}^{dial_i}_{da} + \mathcal{L}^{dial_j}_{da})}{\gamma_2^2} + \log(\gamma_1) + \log(\gamma_2), \tag{11}$$

where $\mathcal{L}^{dial_i}_{da}$ and $\mathcal{L}^{dial_j}_{da}$ are the losses of DAP for dialogues in pair $\phi = (dial_i, dial_j)$, γ_1 and γ_2 are trainable parameters to balance the impact of losses. We compute the gradient of $\mathcal{L}$ to update the parameters of both DiCoh and DAP models.

4 Experiments

4.1 Dialogue Corpora

We compare our approach with several previous dialogue coherence models on DailyDialog (Li et al., 2017) and SwitchBoard (Jurafsky and Shriberg, 1997) as two benchmark English dialogue corpora. Table 2 shows some statistics of these corpora.

DailyDialog contains human-written dialogues about daily topics (e.g. ordinary life, relationships, work, etc) collected by crowd-sourcing. Crowd-workers also annotated utterances with generic DA labels from the set {Inform, Question, Directive, Commissive}. Dialogues in this corpus contain a few utterances (≈ 8) making them more on topic

	DailyDialog	SwitchBoard
# dialogues	13,118	1,155
# DA labels	4	42
avg. # utter. per dialogue	7.9	191.9
avg. # words per utter.	14.6	9.26

Table 2: The statistics of the DailyDialog and SwitchBoard corpora.

and less dispersed. However, utterances are long in terms of the number of words (≈ 15).

SwitchBoard contains informal English dialogues collected from phone conversations between two mutually unknown human participants. The participants were given only one of 70 possible topics as initial topic to start a conversation but they were free to diverge from that topic during the conversation. So, there is no concrete topic associated with each dialogue in this dataset as it is the case for dialogues in DailyDialog.

DA labels in SwitchBoard are about 10 times more fine-grained than those in DailyDialog. For example, a question utterance in SwitchBoard may have a fine-grained DA label such as Yes-No-Question, Wh-Question, Rhetorical-Questions, etc. The distribution of these acts is however highly unbalanced in SwitchBoard: the most frequent act label makes up for 36% of the utterances in the corpus, the three most frequent acts together make up for 68% of the utterances, while most of the remaining act labels just make up for 1% or less of all the utterances.

On average, dialogues in SwitchBoard contain more utterances than those in DailyDialog (192 vs 8) but utterances in SwitchBoard are shorter than those in DailyDialog (9 vs 15). This means that dialogues in SwitchBoard are more likely to span different topics than the ones in DailyDialog. The utterances in DailyDialog are explicitly cleaned of any noise, like "uh-oh", or interruptions by the other speaker, as it is commonly the case for dialogues in SwitchBoard. While each dialogue turn of dialogues in DailyDialog contains only one utterance, dialogue turns in SwitchBoard may consist of several utterances. That is why we consider each dialogue as a sequence of dialogue utterances.

4.2 Problem-domains

The goal of our experiments is to assess if a coherence model assigns coherence scores to dialogues so that a more coherent dialogue obtains a higher score than a less coherent one. Since dialogues in the examined corpora, i.e. DailyDialog and SwitchBoard , are not associated with any coherence assessment score, we synthetically define four perturbation methods to destroy the coherence of dialogues in these corpora, and create a set of dialogue pairs for training and testing coherence models.

We borrow Utterance Ordering (UO) and Utterance Insertion (UI) from previous studies on coherence assessment (Barzilay and Lapata, 2005; Cervone et al., 2018) and also introduce Utterance Replacement (UR), and Even Utterance Ordering (EUO) as more challenging and dialogue-relevant perturbation methods. Since each experiment follows a specific perturbation method, henceforth, we refer to these perturbations as problem-domains:

Utterance Ordering (UO) We randomly permute the order of utterances in dialogue. The original dialogue is preferred over the perturbed one.

Utterance Insertion (UI) We remove each utterance of a dialogue and then re-insert it in any possible utterance position in the dialogue. We assume that the original place of the utterance is the best place for the insertion. Therefore, a coherence model ideally discriminates the original dialogue from the perturbed ones, which are obtained by re-inserting the removed utterance in any utterance position except its original one. This problem-domain is more difficult to solve than UO as the distinction between dialogues is in the position of only one utterance.

Utterance Replacement (UR) We randomly replace one of the utterances in a dialogue with another utterance that is also randomly selected from another dialogue. The original dialogue is preferred over the dialogue generated by UR. Unlike the other problem-domains, which perturb the structure of a dialogue, this problem-domain perturbs the coherence of a dialogue at its semantic level.

Even Utterance Ordering (EUO) This problem-domain is similar to UO but here we re-arrange the order of utterances that are said by one speaker and keep the order of the other utterances, which are said by the other speaker, fixed. Therefore, EUO is more challenging and dialogue-relevant than UO. This problem-domain assesses to what extent coherence models capture the coherence among utterances that are said by one of the speakers in a dialogue.

4.3 Problem-domain Datasets

To create dialogue pairs for each problem-domain, we use the splits provided by the DailyDialog corpus; and for SwitchBoard we take 80% of dialogues for the training, 10% for the validation and 10% for the test sets. Following Cervone et al. (2018), for any dialogue in each set we create 20 perturbations where each of which makes two pairs with the original dialogue. Given dialogue $dial_i$ and its perturbation $dial_j$, we define two dialogue pairs: $(dial_i, dial_j)$ with preference coherence label $l^c = 0$ and $(dial_j, dial_i)$ with label $l^c = 1$.

4.4 In problem-domain Evaluation

In this evaluation, we train, fine-tune, and evaluate our models on the training, validation, and test sets of each problem-domain. Note that these sets are constructed by the same perturbation method.

Compared coherence models We compare the following coherence models in this evaluation: (1) **Random:** This baseline model randomly ranks dialogues in an input dialogue-pair. (2) **CoSim** (Zhang et al., 2018; Xu et al., 2018)**:** This model represents utterances by averaging the pre-trained embeddings of their words. Then, the average of the cosine similarities between vectors of adjacent utterances is taken as the coherence score. In this model, utterance vectors are made using content words by eliminating all stop words. (3) **ASeq** (Gandhe and Traum, 2016)**:** This model relies only DAs transitions and is agnostic to semantic relationships (such as entity transitions) between utterances. Coherence features in this model are the probabilities of n-grams across the sequence of DAs associated with the utterances in dialogue. These features are supplied to a SVM to rank dialogues. (4) **EAGrid** (Cervone et al., 2018)**:** This is the best performing model presented by Cervone et al. (2018) that benefits from both entity and DA transitions between utterances. It represents semantic relationships across utterances via a grid, whose rows are associated with utterances and all columns represent entities but one that represents DAs. Entities are a set of mentions that are extracted by a co-reference system. Entries at the intersections between entity columns and an utterance row represent the grammatical role of an entity in an utterance. The intersection of the DA column and an utterance shows the DA label of the utterance. Cervone et al. (2018) use grammatical role transitions of entities as well as DA label transitions across utterances as indicative patterns for coherence. The frequencies of these patterns are taken as coherence features, which are supplied to Support Vector Machines (SVMs) to discriminate dialogues with respect to their coherence. (5) **S-DiCoh:** This is our coherence model, DiCoh, trained by only the supervision signal for coherence ranking, with the total loss $\mathcal{L} = \mathcal{L}^{\phi}_{coh}$ (see Equation 11). This model does not benefit from DA information to enrich utterance vectors. (6) **M-DiCoh:** This is our full model trained by the proposed MTL using the supervision signals for both coherence ranking and DAP. The main advantage of this model is that it learns to focus on salient information of utterances for coherence assessment based on the given DAs for utterances.

We follow former coherence papers (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013; Mesgar and Strube, 2018; Cervone et al., 2018) and use *accuracy* as the evaluation metric. In our experiments, this metric equals the frequency of correctly discriminated dialogue pairs in the test set of a problem-domain.

$$acc = \frac{\text{\# of correctly discriminated dialogue pairs}}{\text{\# of dialogue pairs}}. \tag{12}$$

To reduce the risk of randomness in our experiments, we run each experiment five times with varying random seeds and report their average (Reimers and Gurevych, 2018).

Settings Each batch consists of 128 and 16 dialogue-pairs for the DailyDialog and SwitchBoard corpora, respectively. Utterances are zero-padded and masked. We use pretrained GloVe embeddings (Pennington et al., 2014) of size 300 wherever word embeddings are required (i.e., in CoSim, S-DiCoh, and M-DiCoh). For the CoSim model, we use the SMART English stop word list (Salton, 1971) to eliminate all stop words. For the ASeq model, we use bi-grams of DA labels to define the coherence features (Cervone et al., 2018). All parameters of the EAGrid model have the same value as the best performing model proposed by Cervone et al. (2018).

In DiCoh, the size of the hidden states in LSTMs of the utterance module is 128 and of the dialogue module is 256. The parameters of this model are optimized using the Adam optimizer where its parameters have default values except the learning rate which is initiated with 0.0005. A dropout layer with $p = 0.1$ is applied to the utterance vectors. We

	DailyDialog				SwitchBoard			
Model	UO	UI	UR	EUO	UO	UI	UR	EUO
Random	50.10	49.97	49.97	49.92	49.98	50.02	49.99	50.13
CoSim	57.20	50.88	65.18	66.86	82.84	55.63	50.87	74.48
ASeq	68.21	57.41	61.89	62.73	**99.70**	73.94	63.48	99.20
EAGrid	71.72	60.93	68.49	67.18	99.65	73.70	**75.61**	**99.83**
S-DiCoh	94.23 ± .74	83.33 ± .81	81.89 ± .26	86.38 ± .29	95.51 ± .61	80.60 ± 1.12	53.61 ± .35	88.83 ± .35
M-DiCoh	**95.92 ± .12**	**88.20 ± .36**	**83.02 ± .50**	**88.55 ± .39**	99.41 ± .11	**85.04 ± 1.14**	58.67 ± 1.79	97.08 ± .20

Table 3: The accuracy (%) of the examined models on the test set of each experiment defined on DailyDialog and SwitchBoard.

train the model for 20 epochs on DailyDialog and 10 epochs on SwitchBoard and evaluate it at the end of each epoch on the validation set. The best performing model on the validation set is used for the final evaluation on the test set. Parameters γ_1 and γ_2 (see Equation 11) are initiated with 2.0 and are updated during training. To have fair comparisons, we train and evaluate all compared models on identical training, validation, and test sets.

Results Table 3 shows the accuracy of the baseline models (top) and our model (bottom) on DailyDialog and SwitchBoard.

We investigate how well our DiCoh model performs in comparison with its baseline peers that do not take DAs into account, i.e., Random and CoSim. We observe that S-DiCoh strongly outperforms these models for all the examined problem-domains on both DailyDialog and SwitchBoard, confirming the validity of our DiCoh model for capturing the semantics of utterances.

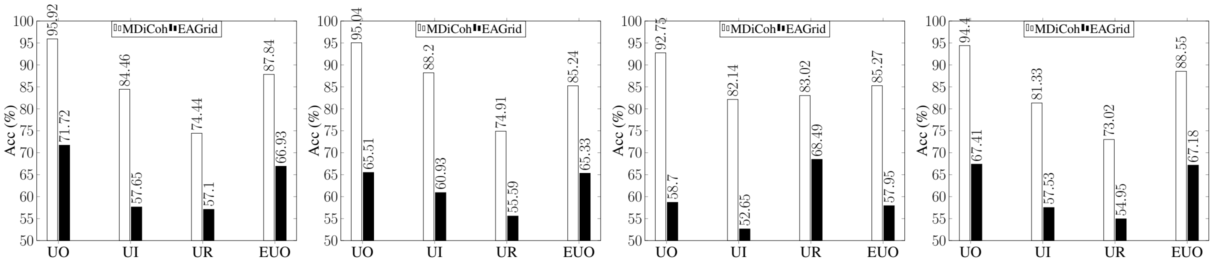
In a more challenging comparison, we compare S-DiCoh with ASeq and EAGrid as the baseline models that use DA information. Our S-DiCoh even surpasses these models for all problem-domains on DailyDialog. However, on SwitchBoard, S-DiCoh achieves lower accuracy than these models for all problem-domains except UI. This observation shows that when dialogue utterances are short (like those in SwitchBoard in comparison with those in DailyDialog), DAs are more crucial for coherence assessment. It is worth noting that unlike EAGrid and ASeq, S-DiCoh is completely agnostic to DA information.

When we employ DAP as an auxiliary task to train the DiCoh model in our MTL setup, we observe that M-DiCoh substantially outperforms the Random, CoSim, and S-DiCoh models (which do not use DAs) for all problem-domains on both DailyDialog and SwitchBoard. It concludes that our proposed MTL approach effectively leverages the DAP task to learn informative utterance vectors for dialogue coherence assessment.

Compared with the ASeq and EAGrid models, which explicitly use gold DA labels during evaluations, our M-DiCoh achieves the highest accuracy for all problem-domains on DailyDialog, showing that our approach for involving DAs yields more informative utterance representations for coherence assessments. However, on SwitchBoard, M-DiCoh increases the accuracy of S-DiCoh up to those of EAGrid for UO and EUO. Surprisingly, it achieves lower accuracy than what EAGrid achieves for UR.

An explanation for why M-DiCoh outperforms ASeq and EAGrid on DailyDialog but not on SwitchBoard might be that the ASeq and EAGrid models explicitly use *gold* DA labels during evaluation but M-DiCoh does not; and the DA labels in SwitchBoard are about 10 times higher fine-grained than those in DailyDialog (see Table 2). This interpretation becomes more concrete by observing a considerable reduction in the performance of ASeq and EAGrid when they are evaluated on DailyDialog compared with when they are evaluated on SwitchBoard. In contrast, our M-DiCoh, which uses DAs only during training to obtain better utterance vectors, performs almost evenly on both corpora. Since our model does not need DA labels during evaluations, it is more suitable than the examined models for evaluating dialogue coherence in real scenarios.

Finally, to shed some light on which parts of a dialogue receive higher attentions by our M-DiCoh model, we analyze the attention weights it assigns to utterance words. Table 4 illustrates the attention weights for an example dialogue from the training set of the UO problem-domain on DailyDialog, where words with higher attention weights are darker than the those with lower attention weights. We observe that using dialog act prediction as an auxiliary task helps our coherence model to assign high attention weights to the salient words in dialogue utterances. The wh-question, adjectives, and



(a) UO (b) UI (c) UR (d) EUO

Figure 3: Comparing EAGrid (black bars) and M-DiCoh (white bars) in cross problem-domain. The labels of figures are the perturbations of the training sets and the labels on x-axes are the perturbations of the test sets.

the verb in questions have higher attention weights; while in other utterances, nouns, e.g. *outlet*, *inexpensive*, *prices*, are more salient. So, our multi-task learning approach yields richer representations of dialog utterances for coherence assessment.

4.5 Cross Problem-domain Evaluation

In a more challenging evaluation setup, we use the model trained on the training set of one problem-domain to evaluate it on the test sets of the other problem-domains. Therefore, the perturbation methods used for constructing the training sets differ from those used for creating the test sets. We compare EAGrid as the state-of-the-art coherence model, and M-DiCoh as our complete model, for cross problem-domain evaluations on DailyDialog.

Results Figure 3 shows the results on the test sets of the problem-domains, where the models are trained on the training set created by the (a) UO, (b) UI, (c) UR, and (d) EUO perturbations. For all perturbations used to construct the training sets, we observe that M-DiCoh outperforms EAGrid for all test perturbations. Interestingly, among all examined perturbations, both M-DiCoh and EAGrid achieve the highest accuracy on UO. We speculate that this perturbation is easy-to-solve as it rearranges all utterances in a dialogue. Cervone et al. (2018) also show that UR is easier to solve than UI.

We note a low-discrepancy in the accuracy of the M-DiCoh model on the test set of UO when the model is trained on the training sets of the different examined problem-domains. The biggest drop in accuracy (3.2 percentage point) on the UO problem-domain is for when the model is trained on the training set of the UR problem-domain. In contrast, we observe a high-discrepancy in the accuracy of the EAGrid model for the UO problem-domain when the model is trained on the training sets of different problem-domains. The accuracy of EAGrid on the test set of UO drops from 71.72% (when trained for UO) to 58.7% (when trained for UR). This is about 13 percentage points drop in accuracy. These results confirm that our M-DiCoh model is more robust than the EAGrid model against different types of perturbation.

	Utterance	DA labels
utt_1	*hello , where can i buy an inexpensive cashmere sweater ?*	Question
utt_2	*maybe you should look around for an outlet .*	Directive
utt_3	*that is a wonderful idea .*	Commisive
utt_4	*outlets have more reasonable prices .*	Inform
utt_5	*thank you for your help .*	Inform
utt_6	*no problem . good luck .*	Inform

Table 4: An illustration of attention weights assigned to words in a dialogue from DailyDialog. Different gray shades show different attention weights.

4.6 DAP Model Evaluation

Since using DAP as an auxiliary task improves the performance of our coherence model; in this experiment, we investigate the impact of MTL on the performance of the DAP model. We train our DAP model without any coherence supervision signal, S-DAP, with $\mathcal{L} = \frac{\mathcal{L}_{da}^{dial_i} + \mathcal{L}_{da}^{dial_j}}{2}$ in Equation 11, and compare it with the model that is trained with our MTL, M-DAP.

Results Table 5 shows the F1 metric[2] of these models for our problem-domains on the DailyDialog dataset. This dataset is larger than SwitchBoard, and the frequency of dialogue act labels in this dataset is more balanced than those in SwitchBoard. We use an SVM classifier supplied with Bag-of-Word representations of utterances as a baseline to put our results in context.

Both S-DAP and M-DAP models outperform the SVM-BoW model for all problem-domains, indi-

[2] We use F1 because there are more than two DA labels.

	UO	UI	UR	EUO
SVM-BoW	76.11	75.52	74.49	75.73
S-DAP	$78.10_{\pm.20}$	$79.15_{\pm.34}$	$77.99_{\pm.35}$	$78.81_{\pm.31}$
M-DAP	$77.32_{\pm.36}$	$78.49_{\pm.33}$	$77.52_{\pm.27}$	$78.51_{\pm.23}$

Table 5: The F1 metric of the DAP model for the test sets of the problem-domains on DailyDialog. S-DAP is the model trained without any coherence supervision, and M-DAP is the model trained with MTL.

cating that the employed DAP model is suitable for solving this task. However, we observe that the M-DAP model works on par with the S-DAP model. This observation shows that the information encoded by the coherence model is not useful for solving the dialogue act prediction task. The coherence model captures semantic relations in a dialogue by encoding information about the content of utterances. Dialogue acts, which indicate speakers' intentions of stating utterances in a dialogue, are independent of the content of utterances, therefore information learned by the coherence model does not help the DAP model.

However, as the other experiments in this paper demonstrate, DAs can help to obtain more informative utterance representations to model dialogue coherence. Our multi-task learning approach relieves the need for explicit DA labels for coherence assessments, which is the main goal of this paper.

5 Conclusions

We propose a novel dialogue coherence model whose utterance encoder layers are shared with a dialogue act prediction model. Unlike previous approaches that utilize these two models in a pipeline, we use them in a multi-task learning scenario where dialogue act prediction is an auxiliary task. Our coherence method outperforms its counterparts for discriminating dialogues from their various perturbations on DailyDialog, and (mostly) performs on par with them on SwitchBoard. Our model (a) benefits from dialogue act prediction task during training to obtain informative utterance vectors, and (b) alleviates the need for gold dialogue act labels during evaluations. These properties holistically make our model suitable for comparing different dialogue agents in terms of coherence and naturalness. For future work, we would like to deeply study the impacts of our perturbations on the coherence of the examined dialogues. We will also investigate to what extent the rankings of dialogues obtained by our model correlate with human-provided rankings.

Acknowledgments

This work was supported by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1). We thank Kevin Stowe and Leonardo Filipe Rodrigues Ribeiro for their valuable feedback on earlier drafts of this paper. We also thank anonymous reviewers for their useful suggestions for improving the quality of the paper.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pages 141–148.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics,* Los Angeles, Cal., 2–4 June 2010, pages 681–684.

Donna K. Byron and Amanda Stent. 1998. A preliminary model of centering in dialog. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, pages 1475–1477.

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association,* Hyderabad, 2–6 September 2018, pages 1011–1015.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, Minnesota., 2–7 June 2019, pages 3806–3812.

Youmna Farag and Helen Yannakoudakis. 2019. Multi-task learning for coherence modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Florence, Italy, 28 July – 2 August, 2019, pages 629–639.

Sudeep Gandhe and David Traum. 2008. Evaluation understudy for dialogue coherence models. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue,* Columbus, Ohio, 19–20 June 2008, pages 172–181.

Sudeep Gandhe and David Traum. 2016. A semi-automated evaluation metric for dialogue model coherence. In *7th International Workshop on Spoken Dialogue Systems,* Saariselk"a, Finland, 13–16 January 2016, pages 141–150.

Yang Gao, Christian M. Meyer, Mohsen Mesgar, and Iryna Gurevych. 2019. Reward learning for efficient reinforcement learning in extractive document summarisation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence,* Macao, China, 10–16 August 2019, pages 2350–2356.

Marjan Ghazvininejad, Chris Brockett, Ming Wei Chang, Bill Dolan, Jianfeng Gao, Wen Tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the 32ed Conference on the Advancement of Artificial Intelligence,* New Orleans, Louisiana, 2–7 February 2018, pages 5110–5117.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Sofia, Bulgaria, 4–9 August 2013, pages 93–103.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. London, U.K.: Longman.

Daniel Jurafsky and Elizabeth Shriberg. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado at Boulder.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition,* Salt Lake City, UT, 18–22 June 2018, pages 7482–7491.

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),* Taipei, Taiwan, 27 November – 1 December, 2017, pages 986—-995.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2019. Dialogue generation: From imitation learning to inverse reinforcement learning. In *Proceedings of the 33rd Conference on the Advancement of Artificial Intelligence,* Honolulu, Hawaii, 21 January –1 February 2019, pages 6722–6729.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* Austin, Texas, 1–5 November 2016, pages 2122–2132.

Mohsen Mesgar and Michael Strube. 2014. Normalized entity graph for computing local coherence. In *Proceedings of TextGraphs-9: Graph-based Methods for Natural Language Processing, Workshop at EMNLP 2014,* Doha, Qatar, 29 October 2014, pages 1–5.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* Brussels, Belgium, 31 October – 4 November 2018, pages 4328–4339.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 1532–1543.

C. Raymond Perrault and James F. Allen. 1978. Speech acts as a basis for understanding dialogue coherence. In *Theoretical Issues in Natural Language Processing-2*.

Amruta Purandare and Diane J. Litman. 2008. Analyzing dialog coherence using transition patterns in lexical and semantic features. In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference,* Coconut Grove, Florida, 15–17 May 2008, pages 195–200.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, Minnesota., 2–7 June 2019, pages 3727–3733.

Nils Reimers and Iryna Gurevych. 2018. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *CoRR*, abs/1803.09578.

Gerard Salton. 1971. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Englewood Cliffs, N.J.: Prentice Hall.

John Searle. 1969. *Speech Acts*. Cambridge University Press, Cambridge, U.K.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the 30th Conference on the Advancement of Artificial Intelligence,* Phoenix, Arizona, 12–17 February 2016, pages 3776–3783.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Vancouver, Canada, 30 July – 4 August 2017, pages 1320–1330.

Svitlana Vakulenko, Maarten de Rijke, Michael Cochez, Vadim Savenkov, and Axel Polleres. 2018. Measuring semantic coherence of a conversation. In *Proceedings of the 17th International Semantic Web Conference,* Monterey, Ca., 8-12 October 2018, pages 634–651.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* Brussels, Belgium, 31 October – 4 November 2018, pages 3981–3991.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing coherence for sequence to sequence model in dialogue generation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence,* Stockholm, Sweden, 13–19 July 2018, pages 4567–4573.

A More Details on EAGrid

EAGrid is a recent model for dialogue coherence with which we compare our models. It mainly extends the entity grid representation for monologue texts. Entity grid is a matrix whose rows represent dialogue utterances and columns encode entities mentioned in dialogue. Each entry in an entity grid is filled by the grammatical role (i.e. subject (“S”), object (“O”), neither of them (“X”)) of its corresponding entity in its corresponding utterance if the entity is mentioned in the utterance, otherwise it is filled by “–”. EAGrid appends a column for encoding dialogue acts to the entity grid such that the entries associated with this column are filled by the dialog act labels of corresponding utterances. Figure 4 shows the EAGrid representation of the example dialogue presented in the top-part of Table 1 in this paper. The grid is generated using EAGrid’s code released by its authors. The probabilities of entities’ grammatical role and dialogue act label transitions of length n across utterance are used as coherence features[3]. These features are supplied to a *SVM* to rank dialogues concerning their coherence.

	Entities				DA labels
	CHARLES	CAPTAIN	UNCLE	THIS	
utt_1	X	–	X	S	inform
utt_2	–	–	–	–	question
utt_3	–	X	–	–	inform
utt_4	–	–	–	–	inform
utt_5	–	–	–	–	inform

Figure 4: The EAGrid representations of Dialogue presented in Table .

B LSTM

As the LSTM layer used in our model is well-known, we give the details of its definition here:

$$\begin{aligned}
i_t &= \sigma(W_{ii}e_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}),\\
f_t &= \sigma(W_{if}e_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}),\\
g_t &= \tanh(W_{ig}e_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}),\\
o_t &= \sigma(W_{io}e_t + b_{io} + W_{ho}h_{(t-1)} + b_{h0}),\\
c_t &= f_t * c_{(t-1)} + i_t * g_t,\\
h_t &= o_t * \tanh(c_t),
\end{aligned} \tag{13}$$

where h_t, is the hidden and c_t is the cell state at word t. The input, forget, cell, and output gates at word t are shown by i_t, f_t, g_t, and o_t, respectively. σ is the Sigmoid function, and $*$ is the Hadamard product. The hidden state is initialized with a zero vector for representing each utterance in dialogue.

C Hyperparameters and Training

To approximate the best values of the hyperparameters, we perform a grid search, in which one parameter is varied while all others are fixed. The search was carried out in the multi-task learning setup on the dataset for the UO problem-domain on DailyDialog. For each variation of hyperparamter values, we train the model on the training set of UO and evaluate it on its validation set. The parameter values that result in the highest respective performance was chosen for evaluation on the test set. The values for the number of epochs and batch size were chosen to trade off the running time and memory consumption of the training. For the experiments on DailyDialog we set the maximum number of epochs to 20 and the batch size to 128, while for SwitchBoard the maximum number of epochs is set to 10 and the batch size to 16. Note that hyperparameter tuning has not been performed for SwitchBoard. Thus for the experiments on SwitchBoard mostly the same hyperparameters as those used for the experiments on DailyDialog are used, with the exception of the batch size and the number of epochs. Table 6 shows the final values for hyperparameters of our models. The optimization is performed by Adam with its default parameter values except for the learning rate. We train the model on the shuffled batches of training data. The model is evaluated on the validation set at each epoch. The model with the best performance on the validation set is chosen for the evaluation on the test set. The training procedure is accelerated by the usage of a Tesla P100 GPU running with CUDA v.10.1, while the model is implemented in the pytorch7 framework version 1.1.0.

parameter	DailyDialog	SwitchBoard
epochs	20	10
batch size	128	16
learning rate	0.0005	0.0005
number of LSTM layers	1	1
hidden layer of LSTM$_u$	128	128
hidden layer of LSTM$_d$	256	256
DA dropout rate	0.1	0.1

Table 6: The values of hyperparameters that result in the best performance on the validation set.

[3] Following the EAGrid model, we set $n = 2$.