Masking Orchestration: Multi-task Pretraining for Multi-role Dialogue Representation Learning

Tianyi Wang*[1], Yating Zhang*[1], Xiaozhong Liu[2], Changlong Sun[1], Qiong Zhang[1]

* [1]Alibaba Group, Hangzhou, Zhejiang, China
[2]Indiana University Bloomington, Bloomington, Indiana, USA
will.wty@alibaba-inc.com, ranran.zyt@alibaba-inc.com
liu237@indiana.edu, changlong.scl@taobao.com, qz.zhang@alibaba-inc.com

Abstract

Multi-role dialogue understanding comprises a wide range of diverse tasks such as question answering, act classification, dialogue summarization etc. While dialogue corpora are abundantly available, labeled data, for specific learning tasks, can be highly scarce and expensive. In this work, we investigate dialogue context representation learning with various types unsupervised pretraining tasks where the training objectives are given naturally according to the nature of the utterance and the structure of the multi-role conversation. Meanwhile, in order to locate essential information for dialogue summarization/extraction, the pretraining process enables external knowledge integration. The proposed fine-tuned pretraining mechanism is comprehensively evaluated via three different dialogue datasets along with a number of downstream dialogue-mining tasks. Result shows that the proposed pretraining mechanism significantly contributes to all the downstream tasks without discrimination to different encoders.

Introduction

Multi-role dialogue mining is a novel topic of critical importance, and it offers powerful potentials for a number of scenarios, e.g, the court debate in civil trial where parties from different camps (plaintiff, defendant, witness, judge etc.) are actively involved, the customer service calls arisen from agent(s) and customer, the business meeting engaged with multi-members. Unfortunately, compared with classical textual data, the labeled multi-role dialogue corpus is scarce and expensive. Unsupervised learning, as a critical alternative, can alleviate this problem, while, based on prior experience(Mikolov et al. 2013a; Devlin et al. 2018; Radford et al. ; Peters et al. 2018), pretraining for complex text data can provide an enhanced content representation for the downstream tasks. In this study, we investigate an innovative problem, multi-role dialogue pretraining for various kinds of NLP tasks.

Indubitably, multi-role dialogue is more complex in its discourse structure and sometimes implicit/ambiguous in its semantics. Two major challenges should be highlighted for

Figure 1: Example Dialogue in Court Debate Dataset

topic. First, different characters may not necessarily share the same vocabulary space, and classical NLP algorithms can hardly consume this difference. Take court debate as an example (see Fig. 1). The judge can be more responsible for investigating the facts and reading the court rules while the other litigants answer the questions from the judge. Moreover, with opposite position, plaintiff and defendant's attitudes, sentiments and descriptions to the same topic can be quite different. The second barrier comes from the interactive nature of the dialogue where single utterance, without dialogue-context, barely contains enough semantics. For instance, as Fig. 1 shows, to accurately represent answer from the defendant, judge's question can be critical and necessary. Thus modeling the relationship among adjacent utterances across various parties is essential for dialogue context representation learning. In addition, given the colloquial dialogue content, the external knowledge-base can play a nontrivial role for context representation learning, e.g., the related law articles and legal knowledge graph can provide important auxiliary semantic information to the target trial debate.

Motivated by such observations, in this paper, we explore dialogue context representation learning through four unsupervised pretraining tasks where the training objectives are given naturally according to the nature of the utterance and the structure of the multi-role conversation. Meanwhile,

to address information implicitly, the proposed method enables the dialogue pretraining via joint learning from external knowledge resource(s). To be specific, our proposed tasks of word prediction, role prediction and utterance generation aim at learning high quality representation by randomly masking and recovering the unit component of the dialogue. The auxiliary task of reference prediction is designed for dialogue domain knowledge contextualization.

Our pretraining mechanism is fine-tuned and evaluated on three different dialogue datasets - Court Debate Dataset, Customer Service Dataset and English Meeting Dataset, with two types of downstream tasks - classification and text generation. In the experiments, we mainly testify the significance of each component in the proposed pretraining mechanism with a delicately designed encoder over court debate corpora in legal domain due to its complexity and high dependence on domain knowledge. To verify the generalizability of the proposed pretraining framework, we conduct evaluation on all downstream tasks over the other two datasets. Result shows that the proposed pretraining mechanism can significantly enhance the performance of all downstream tasks without discrimination to different encoders. Furthermore, we provide a new method to integrate multiple resources during pretraining to enrich the dialogue context.

To the best of our knowledge, this work is the pioneer investigation of multi-role dialogue pretraining with multi-tasks and multi-sources. The contributions of this study are as follows: (1) we delicately define four unsupervised pretraining objectives by masking and recovering the unit component in the dialogue context, and all pretraining tasks show positive effects on improving the testing downstream tasks; (2) we propose an innovative and effective pretraining strategy which can be generalized for different domains and different encoders; (3) In the case of small corpus, which is common for dialogue tasks, the proposed pretraining mechanism can be especially effective for quick convergence; (4) To motivate other scholars to investigate this novel but important problem, we make the experiment dataset publicly available.

Problem Formulation

Let $D = \{U_1, U_2, ..., U_L\}$ denote an arbitrary dialogue, containing L utterances where each utterance U_i is composed of a sequence of l words (namely sentence) $S_i = \{w_{i1}, w_{i2}, ..., w_{il}\}$ and the associated role (of the speaker) r_i. As an optional input, for some datasets, dialogue D can associate with a set of M cited references $F = \{f_1, f_2, ...f_M\}$ (e.g., the name of laws in the legal domain cited by the judge during the trial as shown in Fig. 1). In our pretraining schema, we aim at learning high-quality representation of a dialogue by *masking* and then *recovering* its unit components, i.e., word, role, sentence and reference, as well as leveraging multiple resources, e.g., laws in legal domain for trial dialogue. To be clarified, the definition of important notations in the following sections are illustrated as follows:

- D: a debate dialogue containing L utterances;
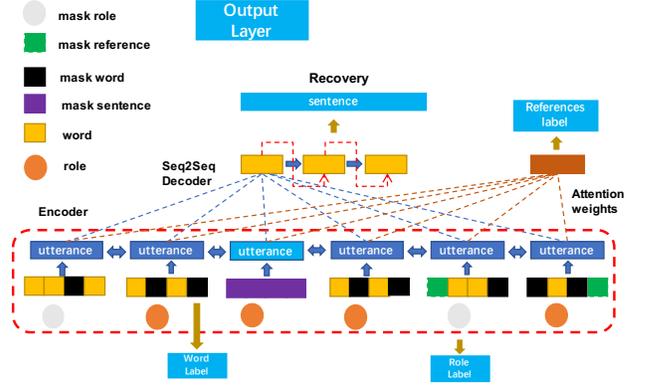- U_i: the i_{th} utterance in D;

Figure 2: Concept Overview of Multi-task Masking Strategy

- r_i: the role of the speaker in U_i (i.e. judge, plaintiff, defendant and witness);
- S_i: the text content of U_i;
- w_{ij}: the j_{th} word in S_i;
- F: the set of cited references in dialogue D (optional);
- f_m: a cited reference in F;
- $\hat{r}_i$: the predicted role of the speaker in U_i;
- $\hat{w}_{ij}$: the predicted j_{th} word in S_i;
- $\hat{S}_i$: the generated text content of U_i;
- $\hat{F} = \{\hat{f}_1, \hat{f}_2, ...\hat{f}_M\}$: the predicted set of cited references in dialogue D (optional)

Note that $\boldsymbol{U_i}$, $\boldsymbol{r_i}$, $\boldsymbol{S_i}$, $\boldsymbol{w_{ij}}$, $\boldsymbol{f_m}$, $\boldsymbol{\hat{r}_i}$, $\boldsymbol{\hat{S}_i}$, $\boldsymbol{\hat{w}_{ij}}$, and $\boldsymbol{\hat{f}_m}$ represent the embedding representations of the corresponding variables in the list.

Multi-task Masking Strategy

Multi-Role Dialogue Encoder

In this section, we first introduce the proposed encoder for delicately representing the hierarchical information in a dialogue, and later in the experiment, we will show that the proposed pretraining mechanism can significantly enhance the performance of downstream tasks without discrimination to different encoders.

• **Utterance Layer** In the utterance layer, we utilize a bidirectional-LSTM to encoder the semantics of the utterance while maintaining its syntactics. To involve the role information into the utterance, we concatenate the role information with each word in the sentence, which is able to project the same word into different dimensional spaces w.r.t. the target role. We hypothesize that the same word may need differentiate when different speakers use it.

$$\mathbf{h_{ij}} = \left[\overrightarrow{\mathbf{LSTM^U}}(\mathbf{e_{ij}}); \overleftarrow{\mathbf{LSTM^U}}(\mathbf{e_{ij}}) \right], j \in [1, l]$$

where $\mathbf{e_{ij}} = [\mathbf{w_{ij}}; \mathbf{r_i}]$. To strengthen the relevance between words in an utterance, we employ the attention mechanism to obtain $\mathbf{U}_i$, which can be interpreted as a local representa-

tion of an utterance:

$$\mathbf{U_i} = \sum_{j=1}^{l} \alpha_j^u \mathbf{h_{ij}}$$

$$\alpha_j^u = \frac{\exp(\mathbf{Q^u h_{ij}})}{\sum_{j'=1}^{l} \exp(\mathbf{Q^u h_{ij'}})}$$

where $\mathbf{Q^u}$ are learnable parameters.

• **Dialogue Layer** To represent the global context in a dialogue, we employ another bidirectional-LSTM to encode the dependency between utterances to obtain a global representation of an utterance, denoted as $\overline{\mathbf{U}}_\mathbf{i}$:

$$\overline{\mathbf{U}}_\mathbf{i} = \left[\overrightarrow{\mathbf{LSTM^D}}(\mathbf{U_i}); \overleftarrow{\mathbf{LSTM^D}}(\mathbf{U_i}) \right], i \in [1, L]$$

$$\overline{\mathbf{U}} = \{\overline{\mathbf{U}}_\mathbf{1}, \overline{\mathbf{U}}_\mathbf{2}, ..., \overline{\mathbf{U}}_\mathbf{L}\} \in \mathbf{R}^{L \times 2dim_h}$$

Then, we feed $\overline{\mathbf{U}}$ to a N-layer Transformer-Block(Vaswani et al. 2017) to suppress the long-term dependency for long dialogue, and finally obtain a dialogue representation $\widetilde{\mathbf{U}} = \mathbf{Transformer^N}(\overline{\mathbf{U}})$, which will be used in the following pretraining tasks as global dialogue context representation.

• **Knowledge Enhance Layer** For some dataset associated with domain knowledge (e.g., in the court debate dataset, the dialogue context can highly rely on the legal domain knowledge, e.g., laws and logic), we propose a Knowledge Enhance Layer to enable external knowledge/resource integration into the utterance representation. At the Knowledge Enhance Layer, the representation of dialogue is enhanced by quoting the content of articles of law, for instance, in the court debate scenario (the masked green parts as shown in Fig. 1). In order to enhance the dialogue representation learning performance from the legal knowledge viewpoint, the proposed dialogue representation is able to learn the vital information from the articles of law in the case context through the attention mechanism. Given a set of cited references F (e.g., all the laws for court debate), we use $C = \{w_1^F, w_2^F, ..., w_t^F\}$ to represent all the content in the references and the corresponding embedding representations is $\mathbf{C} = \{\mathbf{w_1^F}, \mathbf{w_2^F}, ..., \mathbf{w_t^F}\}$. The word embeddings are shared with the words in dialogues. We employ Bi-LSTM to encode the context semantics of C and apply attention mechanism to address the relevance between utterance and the reference:

$$\mathbf{h_n^F} = \left[\overrightarrow{\mathbf{LSTM^F}}(\mathbf{w_n^F}); \overleftarrow{\mathbf{LSTM^F}}(\mathbf{w_n^F}) \right], n \in [1, t]$$

$$\overline{\mathbf{C}}_\mathbf{i} = \sum_{n=1}^{t} \alpha_{in}^c \mathbf{h_n^F}$$

$$\alpha_{in}^c = \frac{\exp(\widetilde{\mathbf{U}}_\mathbf{i}\mathbf{Q^c h_n^F})}{\sum_{n'=1}^{t} \exp(\widetilde{\mathbf{U}}_\mathbf{i}\mathbf{Q^c h_{n'}^F})}$$

where $\mathbf{w_t^F}$ is the t_{th} word in the content of references (e.g., laws) cited in the current dialogue. $\mathbf{Q^c}$ are learnable parameters.

Pretraining with Multi-task Masking Strategy

To host heterogeneous information/knowledge in a dialogue pretraining, in this section, we propose a Multi-task Masking Strategy, which is able to optimize the dialogue representation in terms of four different prediction tasks. The concept overview of the proposed strategy is depicted in Fig. 2.

• **Reference Prediction (F.P.)** Reference prediction is a multi-label classification task for an entire dialogue, which aims at recovering the masked references in a given dialogue. In the experiment, we conduct this masking strategy for court debate dataset where we mask the article names (e.g., Article 8 (4) of the Contract Law) by mining top frequent article names existing in judgment documents. The predicted representation of the references is $\hat{\mathbf{f}} = g(\mathbf{V^f} f_p(\widetilde{\mathbf{U}}) + \mathbf{b^f})$,where $g(*)$ is a non-linear activation function, f_p is a pooling function and $\mathbf{Q^f}, \mathbf{V^f}, \mathbf{b^f}$ are learnable parameters. Finally, we pass $\hat{\mathbf{f}}$ to a fully connected layer and then to a sigmoid function layer for reference prediction. y_k^f denotes the ground truth label and p_k^f is its predicted label. The binary cross-entropy loss function is applied:

$$f_p(\widetilde{\mathbf{U}}) = \sum_{j=1}^{L} \alpha_j^f \widetilde{\mathbf{U}}_\mathbf{j}$$

$$\alpha_j^f = \frac{\exp(\mathbf{Q^f}\widetilde{\mathbf{U}}_\mathbf{j})}{\sum_{i'=1}^{L} \exp(\mathbf{Q^f}\widetilde{\mathbf{U}}_\mathbf{i'})}$$

$$L^f = \sum_{k=1}^{F} \left[y_k^f \log(p_k^f) + (1 - y_k^f) \log(1 - p_k^f) \right]$$

• **Word Prediction (W.P.)** Word prediction is a multiclass classification task. Z denotes the set of all the masked words in dialogue D[1]. For arbitrary sentence S_i and arbitrary masked word w_{ij} in S_i, the predicted representation of the masked word w_{ij} is $\hat{\mathbf{w}}_\mathbf{ij} = g(\mathbf{V^w} \left[\mathbf{h_{ij}}; \widetilde{\mathbf{U}}_\mathbf{i}; \overline{\mathbf{C}}_\mathbf{i} \right] + \mathbf{b^w})$, where $\mathbf{V^w}$ and $\mathbf{b^w}$ are learnable parameters. $\mathbf{h_{ij}}$ is to fetch local contextual information from S_i and $\widetilde{\mathbf{U}}_\mathbf{i}$ is used to enhance the global contextual information from dialogue D, and $\overline{\mathbf{C}}_\mathbf{i}$ helps to involve the external knowledge. Finally, we pass $\hat{\mathbf{w}}_\mathbf{ij}$ to a fully connected layer and then to a softmax function for word prediction. y_{zk}^w denotes the ground truth word $z \in Z$ and p_{zk}^w is the predicted word. The cross-entropy loss function is:

$$L^w = -\sum_{z \in Z} \sum_{k=1}^{N^w} y_{zk}^w \log(p_{zk}^w))$$

• **Role Prediction (R.P.)** Role prediction is also a multiclass classification task. G are the set of all the masked roles in a dialogue[2]. For arbitrary utterance U_i and corresponding masked role r_i, the predicted representation of the masked

---

[1]Base on the prior experience in (Devlin et al. 2018), we randomly mask 15% words for each sentence

[2]In the experiment, we randomly mask the roles of 15% utterances for each dialogue.

Table 1: Statistics of Three Corpus for Pretraining and Downstream Tasks. Note that the #length denotes on average the number of utterances in a dialogue of each corpus.

Corpus	Pretraining			Downstream Tasks	
	#utterance	#dialogue	#length	#utterance	#dialogue
CDD	20M	340K	59	1.6M	6,129
CSD	70M	5M	14	130K	3,463
EMD	1M	32K	31	73K	7,824

Table 2: Pretraining Results over Three Corpus.

Corpus	W.P./acc.	R.P./acc.	S.G./bleu	F.P./acc.
CDD	77.88	84.54	37.30	96.34
CSD	53.82	94.96	18.08	-
EMD	62.63	-	57.74	-

role is $\hat{\mathbf{r}}_\mathbf{i} = g(\mathbf{V^r} \left[\widetilde{\mathbf{U}}_\mathbf{i}; \overline{\mathbf{C}}_\mathbf{i} \right] + \mathbf{b^r})$, where $\mathbf{V^r}$ and $\mathbf{b^r}$ are learnable parameters. Finally, we pass $\hat{\mathbf{r}}_\mathbf{i}$ to a fully connected layer and to a softmax function for role prediction. y^r_{gk} denotes the ground truth role label and p^r_{gk} is the predicted role label. The cross-entropy loss function is:

$$L^r = - \sum_{g \in G} \sum_{k=1}^{N^r} y^r_{gk} \log(p^r_{gk})$$

● **Sentence Generation (S.G.)** Sentence Generation is an NLG task. We utilize encoder-decoder framework with attention mechanism (Luong, Pham, and Manning 2015) for pretraining. We use LSTM cell as basic decoder cell, and $\{[\widetilde{\mathbf{U}}_\mathbf{1}; \overline{\mathbf{C}}_\mathbf{1}], [\widetilde{\mathbf{U}}_\mathbf{2}; \overline{\mathbf{C}}_\mathbf{2}], ..., [\widetilde{\mathbf{U}}_\mathbf{L}; \overline{\mathbf{C}}_\mathbf{L}]\}$ is the encoder representation for masked sentence S_i in dialogue D[3]. The loss function is:

$$L^s = - \log P(S_i \mid D)$$
$$= - \sum_{k=1}^{l} \log P(w_{ik} \mid w_{i1:k-1}, D)$$

The final loss function of the four pretraining objectives is shown as below:

$$L^{total} = L^f + L^w + L^r + L^s$$

which encapsulates various kinds of semantics/knowledge for dialogue pretraining via multi-task masking.

Evaluated Downstream Tasks

To validate the performance and generality of the proposed pretraining mechanism, in this section, we evaluate two types of downstream tasks, *classification* and *summarization*, over three open multi-role dialogue datasets.

Datasets

● **Court Debate Dataset (CDD)** CDD corpus contains over 340K court debate records of civil private loan disputes

cases. The court record is a multi-role debate dialogue associating four roles, i.e., judge, plaintiff, defendant and witness. According to the statistics as shown in Table 1, court debate appears to have longer conversations, on average, containing 59 utterances in a dialogue. We release all the experiment data to motivate other scholars to further investigate this problem[4].

● **Customer Service Dialogue (CSD)** CSD corpus[5] is collected from the customer service center of a top E-commerce platform, which contains over 5 million customer service records between two roles (customer and agent) related to two product categories namely *Clothes* and *Makeup*. The statistics of the dataset is shown in Table 1.

● **English Meeting Dataset (EMD)** EMD corpus is a combined dataset[6] consisting of four open English meeting corpus: AMI-Corpus(Goo and Chen 2018), Switchboard-Corpus(Jurafsky 2000), MRDA-Corpus(Shriberg et al. 2004) and bAbI-Tasks-Corpus[7]. Among the above four corpus, AMI-Corpus includes manually annotated act labels and summaries for the meeting conversations, thus we use such annotated data in AMI-Corpus for downstream tasks. Comparing with the other two datasets, EMD can be much smaller. We use it to validate our hypothesis that the proposed pretraining mechanism can be also efficient for small dialogue corpus.

Classification

Judicial Fact Recognition (**JFR**) is a multi-class classification task, specifically for the court debate corpus in legal domain. The identified judicial facts are the key factors for the judge to analyze and make decision of the target case, thus the objective of this task is to assign each utterance in the court debate to either one of the judicial facts[8] (or the category of *Noise*)[9], to represent the correlation between the utterance and the essential facts. This task mainly evaluates the strength of the pretraining framework on representing the semantics of the complex multi-role debate context as well as differentiating the informative context from the noisy content.

Dialogue Act Recognition (**DAR**) is also a multi-class classification task conducted over the CSD and EMD corpus respectively. The labels in CSD corpus characterize the actions of both customer and staff, i.e., customer side's acts - *advisory*, *request operation*, and staff side's acts - *courtesy reply*, *answer customer's question*, and in total 14 labels are

---

[3]In the experiment, we randomly sample one sentence from dialogue D according to the prior experience of a similar task conducted in (Mehri et al. 2019).

[4]https://github.com/wangtianyiftd/dialogue_pretrain

[5]https://sites.google.com/view/nlp-ssa

[6]For pre-training considerations, relatively large amount of data is required. Thus we combine the four open datasets for pretraining.

[7]https://github.com/NathanDuran/bAbI-Tasks-Corpus

[8]10 fact labels are used in this task: *principal dispute*, *guarantee liability*, *couple debt*, *interest dispute*, *litigation statute dispute*, *fraud loan*, *liquidated damages*, *involving criminal proceedings*, *false lawsuit* and *creditor qualification dispute*.

[9]Statistically, 71.2% of utterances in the experiment data are regarded as noises, i.e., independent to the judicial elements.

involved. Similar in EMD corpus, each utterance is assigned with an act label out of 15 possible labels.

Text Generation

Controversy Focus Generation (**CFG**) is an abstractive summarization task for the court debate corpus. During the civil trial, the presiding judge summarizes the essential controversy focuses[10] according to the plaintiff's complaint[11] and the defendant's answer[12]. Later, the parties from different camps (plaintiff, defendant, witness etc.) debate on court with each other based on the controversy focuses summarized by the presiding judge. This task is challenging in that the construction of abstractive summarization between debated dialogue requires high-quality of global context representation of the entire dialogue which captures the correlation among the utterances by different characters. Thus the pretraining process tends to be significant for initializing parameters of hidden states for the decoders especially when it comes with limited size of training data.

Dialogue Summarization (**DS**) aims at generating a summary for a given dialogue and this task is conducted over EMD corpus. Compared to the text generation task for court debate, the annotated summary in EMD corpus is much shorter and is mainly comprised of key phrases instead of long sentences which describes the topic/intent of a given meeting dialogue[13].

Initialization for Downstream Tasks Training

In the training phase of downstream tasks, for the classification, we use the pretrained dialogue representation (because the dialogue representation has gathered word, sentence and role representation) to initialize the dialogue representation of the classification tasks. The parameters of the decoder part in classification model are randomly initialized (including ones in softmax layer, full connection layer); As for the generation task, we use the pretrained dialogue representation and the parameters of the decoding part (LSTM cell and attention) of the Sentence Generation task (one of the four pretraining tasks), because the decoder structure of the downstream text generation task is the same as that of the pretraining task.

Experimental Settings

Tested Encoders

In the experiment, for each downstream task, we perform pretraining with several state-of-the-art encoders as well as our proposed encoder in this paper. This experiment setting

---

[10]Here shows two examples of the summarized controversy focuses: *Is the loan relationship between the plaintiff and defendant established? Did the plaintiff fulfill he obligation to lend the money?*

[11]The plaintiff's claiming legal rights against another. https://dictionary.law.com/Default.aspx?selected=261

[12]The defendant's pleading to respond to a complaint in a lawsuit. https://dictionary.law.com/Default.aspx?selected=2407

[13]Here shows several examples of the annotated meeting summaries: *evaluation of project process*, *possible issues with project goals*, *closing*, *discussion*, *marketing strategy*.

ensures a universality validation of the proposed pretraining mechanism over a variety of encoding strategies.

For the *classification* tasks, besides our proposed encoder, we also select two models - **HBLSTM-CRF**(Kumar et al. 2018) and **CRF-ASN**(Chen et al. 2018) - as encoders for both pretraining and downstreaming stages. The two models have leading performance on MRDA-Corpus and Switchboard-Corpus as shown on the Leaderboard[14].

As for the *text generation* tasks in CDD and EMD corpus, except for using our own encoder, we use Discourse Aware Hierarchical Sequence-to-Sequence model (**DAHS2S**)(Cohan et al. 2018) employed in (Goo and Chen 2018) as the other encoder for abstractive summarization.

Evaluation Metrics

For *classification* task, we evaluate the performance of each model based on two popular classification metrics: micro F1 and macro F1 scores. To automatically assess the quality of *generated text*, we used ROUGE (Lin and Hovy 2003) and BLEU(Papineni et al. 2002) scores to compare different models. We report ROUGE-1 and ROUGE-2 as the means of assessing informativeness and ROUGE-L as well as BLEU-4 for assessing fluency.

Hyper-Parameter Selection

In our experiments, we optimize the tested models using Adam Optimization(Kingma and Ba 2014) with learning rate of $5e$-4. The dimensions of word embedding and role embedding are 300 and 100 respectively. The size of hidden layers are all set to 256. We use 2 layer Transformer-Block, where feed-forward filter size is 1024, and the number of heads equals to 4.

Results and Discussion

Overall Performance

To evaluate the performance of the proposed pretraining model, we export the results of pretraining tasks as well as the improved performance on downstream tasks over three different datasets. Table 2 shows the performance of the proposed pretraining tasks on all datasets[15] which indicates how effective the proposed pretraining mechanism on recovering the information in the dialogue. According to the pretraining scores, we can also observe the complexity of the corresponding corpus. For instance, The performance of word prediction and sentence generation tasks in CSD corpus is worse than that in CDD corpus due to the word diversity in customer service for coping with a variety of disputes, however in relatively close legal domain, the words of different roles especially of the judges during trial remain similar across different cases. As for the role prediction, in customer service, usually only two characters are involved while in court debate it is common to have multiple roles therefore it might be the reason why the task of role prediction in CDD is lower than that in CSD.

---

[14]http://nlpprogress.com/english/dialogue.html

[15]Note that there is no reference used in CSD and EMD corpus and EMD corpus contains no character information neither, so the corresponding pretraining tasks are omitted for the two corpus.

Table 3: Downstream Task - Classification Results with Different Encoders over Three Corpus. † at "pretrain" rows indicates statistically significant difference from the corresponding value of "vanilla" model ($p < 0.01$).

Model	CDD: JFR		CSD: DAR		EMD: DAR	
	micro F_1	macro F_1	micro F_1	macro F_1	micro F_1	macro F_1
HBLSTM-CRF(vanilla)	81.60	34.14	84.96	76.03	66.04	51.53
HBLSTM-CRF(pretrain)	**81.68**†	**39.13**†	**85.34**†	**77.36**†	**66.22**	**51.94**
CRF-ASN(vanilla)	80.90	31.15	83.55	73.75	64.13	44.12
CRF-ASN(pretrain)	**81.55**†	**38.22**†	**83.92**†	**74.93**†	**64.58**†	**49.55**†
Our model(vanilla)	81.73	39.61	85.34	76.83	66.83	52.78
Our model(pretrain)	**82.06**†	**45.45**†	**85.80**†	**78.28**†	**67.34**†	**53.69**†

Table 4: Downstream Task - Summarization Results with Different Encoders over Two Corpus. ⋆ at "pretrain" rows indicates statistically significant difference from the corresponding value of "vanilla" model ($p < 0.001$).

| Model	CDD: CFG					EMD: DS				
	rouge-1	rouge-2	rouge-3	rouge-L	bleu4	rouge-1	rouge-2	rouge-3	rouge-L	bleu4
DAHS2S(vanilla)	26.83	7.27	4.15	22.21	5.27	35.43	28.84	25.95	34.33	20.69
DAHS2S(pretrain)	**34.94**⋆	**12.98**⋆	**7.18**⋆	**29.51**⋆	**8.06**⋆	**40.68**⋆	**32.11**⋆	**27.00**⋆	**39.60**⋆	**22.26**⋆
Our model(vanilla)	22.55	3.99	1.66	18.95	2.94	31.00	26.40	25.23	29.83	14.87
Our model(pretrain)	**36.55**⋆	**13.54**⋆	**7.48**⋆	**30.84**⋆	**8.59**⋆	**41.39**⋆	**34.18**⋆	**29.64**⋆	**40.34**⋆	**23.20**⋆
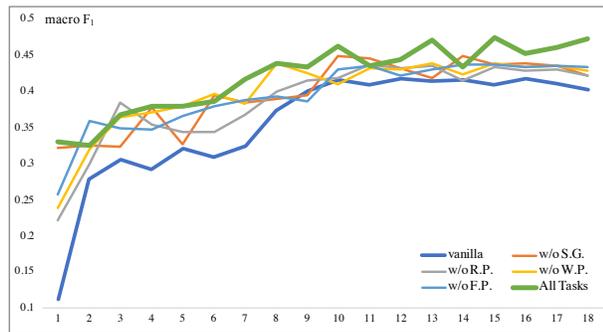
(a) Task JFR (macro F_1 score)
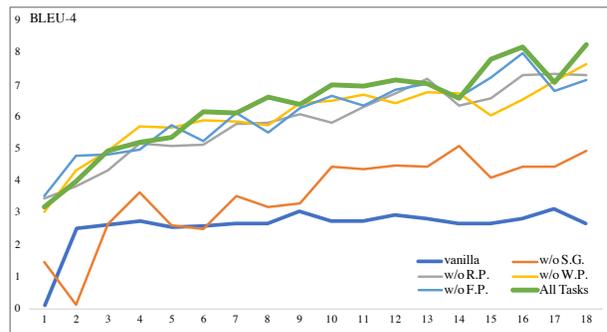
(b) Task CFG (BLEU-4 score)

Figure 3: The performance of two downstream tasks with different pretraining objectives.

Table 3 and 4 demonstrate the performance on two downstream tasks respectively over three datasets where the notation "vanilla" represents the randomly initialized encoders used for the downstream tasks while "pretrain" denotes the pretrained encoders for the downstream tasks.

● **Classification** Table 3 aggregates the classification results on all three datasets. In general, we can observe that, for all tested encoders, the pretraining process results are significantly superior than baselines over all datasets under almost all metrics, especially for macro F_1 score. Statistically, on average, pretraining under different encoders achieves 17.4%, 1.4% and 5% increase in macro F_1 score over CDD, CSD and EMD corpus respectively, which implies such pretraining can be very helpful to alleviate the problem of unbalanced/bias data where there are some categories with rather small/large training data (e.g., in CDD corpus, the category of *Noise* takes up more than 70% of utterances.). Moreover, our proposed encoder in this paper outperforms the state-of-the-art encoders for the tested corpus.

● **Text Generation** Table 4 depicts the text generation results on CDD and EMD corpus respectively. Similar to classification task, the pretraining conducted under both encoders shows positive effects on all the evaluation metrics. An interesting finding is that the proposed model achieves limited performance in "vanilla" setting but after pretraining it quickly surpassed the method DAHS2S in all metrics. In addition, as we aforementioned, the **CFG** task is much challenging compared to the **DS** task due to the relatively long text (i.e., controversy focus) needs to be generated via CDD corpus. In such difficult case (CDD corpus), the proposed pretraining method performs beyond expectation. While it successfully estimates the comprehensive dialogue context representation, comparing with the baselines, Bleu-4 scores increased by 53% and 192%. As for the relatively small dataset, EMD corpus, the pretraining method brings about 7.6% and 56% increase for the two tested encoders.

Ablation Test

To assess the contribution of different components in the proposed method, we conduct ablation tests for both clas-

Table 5: Ablation Test on **JFR** Tasks with Different Encoders over Court Debate Dataset

Method	HBLSTM-CRF		ASN-CRF		Our Model	
	miF_1	maF_1	miF_1	maF_1	miF_1	maF_1
All	**81.68**	**39.13**	**81.55**	**38.22**	**82.06**	**45.45**
w/o W.P.	81.63	38.03	81.15	34.17	81.77	41.66
w/o R.P.	81.64	38.38	81.06	32.46	82.01	40.78
w/o S.G.	81.66	38.13	81.27	32.93	81.97	43.81
w/o F.P.	81.65	38.41	81.47	35.07	81.73	41.71

sification (see Table 5) and text generation (see Table 6) tasks on CDD corpus[16]. To prove the generalizability of the proposed pretraining schema on different encoders, for all tested encoders, the same ablation test is conducted by removing each pretraining objective.

Table 5 reports the F_1 scores of **JFR** task, for each encoder, when training on all objectives and when training on all objectives except the particular one. As Table 5 shows, all the model components contribute positively to the results. To be specific, the pretraining task of word prediction has largest impact on the method HBLSTM-CRF. Their removal causes 1.8% relative increase in error (RIE) for macro F_1 scores, while the task of role prediction has biggest impact on the model ASN-CRF (9.3% RIE for macro F_1 score). As for our encoder, reference and word prediction show greatest impact on the performance. In general, we notice that, for classification, the three prediction tasks affect the model effect to varying degrees.

As for **CFG** task, the findings are quite different as suggested in Table 6. Since **CFG** is a text generation task, we can observe that the pretraining task, sentence generation, tends to have largest impact on both tested encoders evaluated by Bleu-4 score. Such observations indicate that the pretraining tasks have strong impact on the downstream tasks in similar types.

Table 6: Ablation Test on **CFG** Tasks with Different Encoders over Court Debate Dataset

Method	rouge-1	rouge-2	rouge-3	rouge-L	bleu4
DAHS2S	34.94	**12.98**	**7.18**	29.51	**8.06**
w/o W.P.	34.65	11.98	6.64	28.75	7.92
w/o R.P.	34.26	12.13	6.39	28.54	7.42
w/o S.G.	30.93	9.64	5.38	25.74	6.40
w/o F.P.	**35.54**	12.69	6.82	**29.77**	7.92
Our Model	**36.55**	**13.54**	**7.48**	**30.84**	**8.59**
w/o W.P.	35.79	13.13	7.10	29.96	8.09
w/o R.P.	35.58	12.59	7.10	30.06	8.22
w/o S.G.	30.38	8.59	4.25	24.85	5.58
w/o F.P.	36.49	13.13	7.21	30.64	8.31

Convergence Analysis

To further validate the performance of the proposed pretraining model, we conduct experiments to monitor the impact of pretraining on the convergence of all downstream tasks. In the experiment, we employ the proposed model as encoder and evaluate the performance of two downstream tasks with

---

[16]Only CDD corpus enables testing on all four pretraining objectives (see Table 2).

pretraining on all objectives and on all objectives except the particular one at each epoch. Fig. 3a and 3b depict the results of **JFR** and **CFG** tasks respectively.

As shown in Fig.3a and 3b, we can observe that the performance of the model with pretraining on all objectives is significantly superior than the "vanilla" one from the initial epoch which indicates the advantage of learning with pretrained parameters instead of random initialization. Compared the pretrained "all tasks" model with the models removing a particular task, the former performs more stably and almost always outperforms the latter ones.

Related Work

Unsupervised Pretraining in NLP

Unsupervised pretraining for natural language becomes popular and widely adopted in many NLP tasks due to the nature that labeled data for specific learning tasks can be highly scarce and expensive. Due to such motivation, the earliest approaches used unlabeled data to compute word-level or phrase-level embeddings(Collobert et al. 2011; Mikolov et al. 2013b; 2013a; Pennington, Socher, and Manning 2014), which were later used as atom features in a supervised model for the further specific learning tasks. Although the pretrained word/phrase-level embeddings could improve the performance on various tasks, such approaches can only capture the atom-level information regardless of the higher-level semantics and syntactics. Recent research work have focused on learning sentence-level and document-level representations which are learned from unlabeled corpus (Radford et al. ; Peters et al. 2018; Chang et al. 2019).

Compared to the sentence or document-level representation learning, dialogue representation learning can be more complex due to its hierarchical structures as well as its heterogenous information resources. In this work, we address the difficulty of such challenges and propose a masking strategy for pretraining in multi-task schema.

Dialogue Representation Learning

Recent research work has focused on proving the effectiveness of hierarchical modeling in dialogue scenarios (Weston 2016). The common approach is focusing on constructing delicate encoders for representing dialogue structures. For instance, Weston(Weston 2016) employed a memory network based encoder to capture the context information in a dialogue for specific task learning. Although there have been plenty of research work focusing on document representation learning, pretraining methods are still in their infancy in the domain of dialogue. Mehri et al(Mehri et al. 2019) recently approaches to such problem by proposing pretraining objectives for dialogue context representation learning. Compared to them, we are different in several aspects: Firstly, to the best of our knowledge, we are the first to involve role information in this area, and in our framework, we are flexible to involve external resources during pretraining; Secondly, all tasks in our work are in bidirectional approach which means we can consider the context in both directions, similar to the strategy to BERT(Devlin et al.

2018); Third, the experimental results demonstrate the generalizability of our proposed pretraining strategy over different domains and along with various types of encoders.

Conclusion

This paper investigates the research problem of dialogue context representation learning by proposing a multi-task masking strategy to perform various types unsupervised pretraining tasks, including word prediction, role prediction, sentence generation and reference prediction. The proposed fine-tuned pretraining mechanism is comprehensively evaluated through three different dialogue datasets along with a number of downstream dialogue-mining tasks. Result shows that the proposed pretraining mechanism significantly contributes to all the downstream tasks without discrimination to different encoders.

Acknowledgments

References

Chang, M.-W.; Toutanova, K.; Lee, K.; and Devlin, J. 2019. Language model pre-training for hierarchical document representations. *arXiv preprint arXiv:1901.09128*.

Chen, Z.; Yang, R.; Zhao, Z.; Cai, D.; and He, X. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 225–234. ACM.

Cohan, A.; Dernoncourt, F.; Kim, D. S.; Bui, T.; Kim, S.; Chang, W.; and Goharian, N. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12(Aug):2493–2537.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Goo, C.-W., and Chen, Y.-N. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 735–742. IEEE.

Jurafsky, D. 2000. *Speech & language processing*. Pearson Education India.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, H.; Agarwal, A.; Dasgupta, R.; and Joshi, S. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lin, C.-Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mehri, S.; Razumovsakaia, E.; Zhao, T.; and Eskenazi, M. 2019. Pretraining methods for dialog context representation learning. *arXiv preprint arXiv:1906.00414*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. Improving language understanding by generative pretraining.

Shriberg, E.; Dhillon, R.; Bhagat, S.; Ang, J.; and Carvey, H. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 97–100.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Weston, J. E. 2016. Dialog-based language learning. In *Advances in Neural Information Processing Systems*, 829–837.