ZERO-SHOT CROSS-LINGUAL CONVERSATIONAL SEMANTIC ROLE LABELING

Anonymous authors
Paper under double-blind review

ABSTRACT

While conversational semantic role labeling (CSRL) has shown its usefulness on Chinese conversational tasks, it is still under-explored in non-Chinese languages due to the lack of multilingual CSRL annotations for the parser training. To avoid expensive data collection and error-propagation of translation-based methods, we present a simple but effective approach to perform zero-shot cross-lingual CSRL. Our model implicitly learns language-agnostic, conversational structure-aware and semantically rich representations with the hierarchical encoders and elaborately designed pre-training objectives. Through comprehensive experiments, we find that, our cross-lingual model not only outperforms baselines by large margins but it is also robust to low-resource scenarios. More impressively, we attempt to use CSRL information to help downstream English conversational tasks, including question-in-context rewriting and multi-turn dialogue response generation. Although we have obtained competitive performance on these tasks without CSRL information, substantial improvements are further achieved after introducing CSRL information, which indicates the effectiveness of our cross-lingual CSRL model and the usefulness of CSRL to English dialogue tasks.

1 INTRODUCTION

Conversational Semantic Role Labeling (CSRL) (Xu et al., 2021) is a recently proposed dialogue understanding task, which aims to extract predicate-argument pairs from the entire conversation. By recovering dropped and referred components in conversation, CSRL has shown its usefulness to a set of Chinese conversation-based tasks, including multi-turn dialogue rewriting (Su et al., 2019) and response generation (Wu et al., 2019). However, there remains a paucity of evidence on its effectiveness towards non-Chinese languages owing to the lack of multilingual CSRL models. To adapt a model into new languages, previous solutions can be divided into three categories: 1) manually annotating a new dataset in the target language (Daza & Frank, 2020) 2) borrowing machine translation and word alignment techniques to transfer the dataset in source language into target language (Daza & Frank, 2019; Fei et al., 2020a) 3) zero-shot transfer learning with multilingual pre-trained language model (Rijhwani et al., 2019; Sherborne & Lapata, 2021). Due to the fact that manually collecting annotations is costly and translation-based methods might introduce translation or word alignment errors, zero-shot cross-lingual transfer learning is more practical to the NLP community.

Recent works have witnessed prominent performances of multilingual pre-trained language models (PrLMs) (Devlin et al., 2019; Conneau & Lample, 2019; Conneau et al., 2020) on cross-lingual tasks, including machine translation (Lin et al., 2020; Liu et al., 2020; Fan et al., 2021), semantic role labeling (SRL) (Conia & Navigli, 2020; Conia et al., 2021) and semantic parsing (Fei et al., 2020b; Sherborne et al., 2020; Sherborne & Lapata, 2021). However, cross-lingual CSRL, as a combination of three challenging tasks (i.e., cross-lingual task, dialogue task and SRL task), suffers three outstanding difficulties: 1) **latent space alignment** - how to map word representations of different languages into an overlapping space; 2) **conversation structure encoding** - how to capture high-level dialogue features such as speaker dependency and temporal dependency; and 3) **semantic arguments identification** - how to highlight the relations between the predicate and its arguments, wherein PrLMs can only encode multilingual inputs to an overlapping vector space in a certain extend. Although there are also some success that can separately achieve structural conversation encoding (Mehri et al., 2019; Xu & Zhao, 2021; Zhang & Zhao, 2021) and semantic arguments identification (Wu et al., 2021; Conia et al., 2021), a unified method for jointly solving these problems is still under-explored, especially in cross-lingual scenario.

In this work, we propose a simple yet effective model to perform zero-shot cross-lingual CSRL. Specifically, our model consists of three modules, namely cross-lingual language model (CLM), structure-aware conversation encoder (SA-Encoder) and predicate-argument encoder (PA-Encoder). First, we use the CLM to map the representations of different languages into an overlapping latent space. Secondly, we feed word embeddings along with temporal and speaker embeddings into SA-Encoder, and obtain conversational structure-aware context representations. Finally, we feed the resulted context representations with predicate vectors into PA-Encoder to classify the semantic roles. In addition, we also propose a hierarchical pre-training method to boost the cross-lingual CSRL performance. Experimental results show that our model is not only superior to all baselines, but also robust to low-resource scenarios. Further experiments of applying CSRL parsing results to help downstream dialogue tasks consistently confirms the usefulness of CSRL to non-Chinese dialogue tasks. We will release our code and checkpoints of our best models at `https://github.com` upon the acceptance.

2 Related Work

Zero-shot cross-lingual transfer learning. Recently, thanks to the rapid development of multilingual pre-trained language models such as multilingual BERT (Devlin et al., 2019), XLM (Conneau & Lample, 2019) and XLM-R (Conneau et al., 2020), a number of approaches have been proposed for zero-shot cross-lingual transfer learning on various downstream tasks, including semantic parsing (Sherborne & Lapata, 2021), headline generation (Shen et al., 2018) and natural language understanding (Liu et al., 2019; Lauscher et al., 2020). In this work, we claim our method is zero-shot because no non-Chinese CSRL annotations are seen during the training stage. For decoding, we directly use the cross-lingual CSRL model trained on Chinese CSRL data to analyze conversations in other languages. In addition, to the best of our knowledge, we are the first one to jointly model conversational and semantic features in zero-shot cross-lingual scenario.

Conversational semantic role labeling. While ellipsis and anaphora frequently occur in dialogues, Xu et al. (2021) observed that most of dropped or referred components can be found in dialogue histories. Following this observation, they proposed conversational semantic role labeling (CSRL) which required the model to find predicate-argument structures over the entire conversation instead of a single sentence. In this way, when analyzing a predicate in the latest utterance, a CSRL model needs to consider both the current turn and previous turns to search potential arguments, and thus might recover the omitted components. Furthermore, Xu et al. (2020; 2021) also confirmed the usefulness of CSRL to dialogue tasks by applying CSRL information into downstream dialogue tasks. However, there are still two main problems to be solved for CSRL task: (1) the performance of current state-of-the-art CSRL model (Xu et al., 2021) is still far from satisfactory due to the lack of high-level conversational and semantic features modeling; (2) the usefulness of CSRL to conversational tasks in non-Chinese has not been confirmed yet due to the lack of multilingual CSRL models. In this work, we primarily focus on the later problem and propose a simple but effective model to perform cross-lingual CSRL. But interestingly, we also find that our cross-lingual CSRL model outperforms all existing models, which further indicats the effectiveness of our method.

3 Methodology

We describe the model architecture at Section 3.1 and the pre-training objectives at Section 3.2.

3.1 Architecture

Cross-lingual Language Model (CLM) Given a dialogue $C = \{u_1, u_2, ..., u_N\}$ of N utterances, where $u_i = \{w_1^i, w_2^i, ..., w_{|u_i|}^i\}$ consisting of a sequence of words, we first concatenate utterances into a sequence and then use a pre-trained cross-lingual language model such as XLM-R (Conneau et al., 2020) or mBERT (Devlin et al., 2019) to capture the syntactic and semantic characteristics. Following Conia et al. (2021), we obtain word representations **e** by concatenating the hidden states of the four top-most layers of the language model.

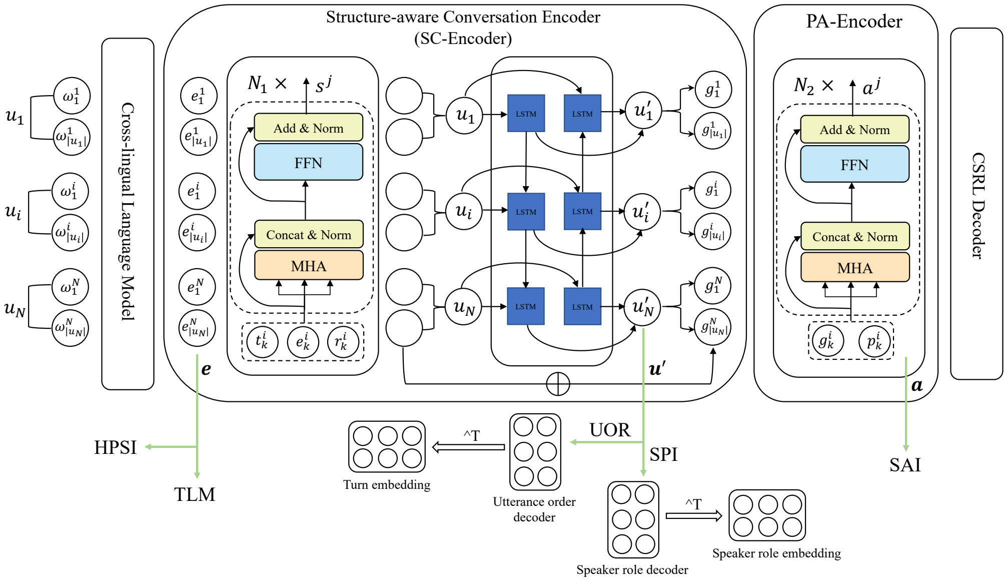

Figure 1: Overall model architecture.

Structure-aware Conversation Encoder (SC-Encoder) Different from standard SRL(Carreras & Màrquez, 2005), CSRL requires model to find arguments from no only the current turn, but also previous turns, leading to more challenges of modeling the dialogue context. To address this problem, we propose a universal structure-aware conversation encoder which comprises of two parts, i.e., word-level encoder and utterance-level encoder. Following Xu et al. (2021), we also incorporate speaker role and dialogue turn indicators to reserve high-level structural features of the dialogue, which could help the model to better handle coreference resolution and zero pronoun resolution. Formally, given a sequence of word representations $\boldsymbol{e} = (\boldsymbol{e}_1^1, ..., \boldsymbol{e}_k^i, ..., \boldsymbol{e}_{|u_N|}^N)$, dialogue turn embeddings $\boldsymbol{t} = (\boldsymbol{t}_1^1, ..., \boldsymbol{t}_k^i, ..., \boldsymbol{t}_{|u_N|}^N)$ and speaker role embeddings $\boldsymbol{r} = (\boldsymbol{r}_1^1, ..., \boldsymbol{r}_k^i, ..., \boldsymbol{r}_{|u_N|}^N)$, the word-level encoder computes a sequence of timestep encodings $\boldsymbol{s}$ as follows:

$$\boldsymbol{s}_{(i,k)}^j = \begin{cases} \boldsymbol{e}_k^i \oplus \boldsymbol{t}_k^i \oplus \boldsymbol{r}_k^i & if \;\; j = 0 \\ \boldsymbol{s}_{(i,k)}^{j-1} \oplus \text{MT}{\small\text{RANS}}^j(\boldsymbol{s}_{(i,k)}^{j-1}) & otherwise \end{cases} \tag{1}$$

where $\boldsymbol{s}_{(i,k)}^j$ is the timestep encoding of k-th tokens in i-th utterance from j-th word-level encoder layer, and MTRANS is the **M**odified **Trans**former encoder layer. Concretely, we drop the `[Add]` operation in the first residual connection layer and replace it with `[Concat]` because we argue that concatenation is a superior approach to reserve the information from previous layers[1].

We obtain utterance representations $\boldsymbol{u}$ by max-pooling over words in the same utterance. Then we pass the resulted utterance representations $\boldsymbol{u}$ through a stack of Bi-LSTM (Hochreiter & Schmidhuber, 1997) layers to obtain the sequentially encoded utterance representations $\boldsymbol{u}'$. Finally, we incorporate $\boldsymbol{u}'$ with context representations $\boldsymbol{s}$ from previous layer to obtain structure-aware dialogue context representations $\boldsymbol{g}$ as follows:

$$\boldsymbol{g}_k^i = \sigma(\mathbf{W}^g[\boldsymbol{s}_k^i \oplus \boldsymbol{u}_i'] + \mathbf{b}^g) \tag{2}$$

where σ is activation function, $\boldsymbol{s}_k^i$ is the encoding of k-th token in i-th utterance from the last layer of the word-level encoder, and $\mathbf{W}^g$ and $\mathbf{b}^g$ are trainable parameters.

Predicate-Argument Encoder (PA-Encoder) We introduce the third module (i.e., predicate-argument encoder) whose goal is to capture the relations between each predicate-argument cou-

[1] We observe slight performance improvements with MTRANS against standard Transformer encoder layer.

ple that appears in the conversation. Similar with the word-level encoder, we use a stack of MTRANS layers to implement this encoder. Formally, denote predicate embedding as $\boldsymbol{p} = (\boldsymbol{p}_1^1, ..., \boldsymbol{p}_k^i, ..., \boldsymbol{p}_{|u_N|}^N)$, the model calculates the predicate-specific argument encoding as follows:

$$\boldsymbol{a}_{(i,k)}^j = \begin{cases} \boldsymbol{g}_k^i \oplus \boldsymbol{p}_k^i & if \;\; j = 0 \\ \boldsymbol{a}_{(i,k)}^{j-1} \oplus \text{MTRANS}^j(\boldsymbol{a}_{(i,k)}^{j-1}) & otherwise \end{cases} \tag{3}$$

where $\boldsymbol{g}_k^i$ is the token embedding from conversation encoder and $\boldsymbol{p}_k^i$ is the corresponding predicate indicator embedding. Finally, we obtain the semantic role encoding $\boldsymbol{l}$ using the resulted argument encodings from the last layer of the predicate-argument encoder:

$$\boldsymbol{l}_k^i = \sigma(\mathbf{W}^l \boldsymbol{a}_k^i + \mathbf{b}^l) \tag{4}$$

In particular, we emphasize that our proposed model is language-agnostic since we do not introduce any language-specific knowledge such as word order, part-of-speech tags or dependent relations, all of which may differ from language to language.

3.2 Pre-training Objectives

Besides the universal model, we also elaborately design five pre-training objectives to model task-specific but language-agnostic features for better cross-lingual performance. In this section, we divide our pre-training objectives into three groups according to the challenges to be solved.

Latent space alignment In cross-lingual language module, we use mBERT or XLM-R to align the latent space of different languages. Although mBERT and XLM-R have exhibited good alignment ability, even both of which are trained with unpaired data, we may further improve it when we have access to parallel data.

Following (Conneau & Lample, 2019), we first use translation language model (TLM) to make direct connections between parallel sentences. Concretely, we concatenate parallel sentences as a single consecutive token sequence with special tokens separating them and then perform masked language model (MLM) (Devlin et al., 2019) on the concatenated sequence. Compared with MLM, TLM objective encourage the model to align the representations of source and target languages. Different from Conneau & Lample (2019), we feed source and target sentences twice in different orders instead of resetting the positions of target sentences.

Besides improving word-level alignment ability by TLM, we also propose to enhance sentence-level alignment ability using hard parallel sentence identification (HPSI). Specifically, we select a pair of parallel or non-parallel sentences from the training set with equal probability. Then the model is required to predict whether the sampled sentence pair is parallel or not . Different from the standard PSI (Dou & Neubig, 2021), we sample the non-parallel sentence upon the n-gram similarity or construct it by text perturbation[2] instead of in a random manner. We think that the closer the negative sample is to the positive sample, the better representations the model can learn.

In practice, we use the initial context representation $\boldsymbol{e}$ from CLM as the input of TLM and HPSI decoders, and pre-train the CLM using the combination of TLM and HPSI, finally achieving latent space alignment.

Conversation structure encoding Although there are a number of pre-training objectives proposed to learn dialogue context representations (Mehri et al., 2019), structural representations (Zhang & Zhao, 2021; Gu et al., 2021) and semantic representations (Wu et al., 2021), we tend to explicitly model speaker dependency and temporal dependency in the conversation following Xu et al. (2021) which incorporates dialogue turn and speaker role information into the model and ultimately obtains good performance.

We first propose speaker role identification (SPI) to learn speaker dependency in the conversation. Specifically, we randomly sample $K_1\%$ utterances and replace their speaker indicators with special

[2] Details in Appendix A

mask tags. To make the task harder and effective, we split the utterances into clauses if only two interlocutors utter in turn in a conversation. Therefore, the goal of SPI is to predict the masked speaker roles according to the given speaker information and context. Secondly, we borrow utterance order permutation (UOR) (Zhang & Zhao, 2021) to encourage the model to be aware of temporal connections among utterances in the context. Concretely, given a set of utterances, we randomly shuffle the last $K_2\%$ utterances and require the model to organize them into a coherent context.

We use the sequentially informed utterance representations $\boldsymbol{u}'$ as the input of speaker role and utterance order decoders, and pre-train the SC-Encoder using the combination of SPI and UOR. After pre-training, we employ the transposed speaker role decoder and utterance order decoder as the speaker role embedding and dialogue turn embedding in CSRL model.

Semantic arguments identification The core of all SRL-related tasks is to recognize the predicate-argument pairs from the input. Therefore, we propose semantic arguments identification (SAI) objective to strength the correlations between the predicate and its arguments using external standard SRL corpus such as CoNLL-2012 in the pre-training stage. Specifically, for each standard SRL sample, we only reserve the spans of the overlapped semantic roles between standard SRL and CSRL, including ARG0-4, ARG-LOC, ARG-TMP and ARG-PRP. Then the model is required to find these textual spans with the given predicate. We think this objective would benefit to boundary detection, especially for location and temporal arguments.

In practice, we drop the utterance-level encoder of SC-Encoder to fit in standard SRL samples since they do not have any conversational characteristics. We directly feed the word-level context representations $\boldsymbol{s}$ into PA-Encoder, and then use argument encodings $\boldsymbol{a}$ to make the classification.

3.3 Training

Hierarchical Pre-training The pre-training is hierarchically conducted according to different modules, and the pre-training of the upper module is based on the pre-trained lower modules. Specifically, we first train CLM module with TLM and HPSI; then we train SC-Encoder with SPI and UOR while keeping the weights of pre-trained CLM module unchanged; finally we train PA-Encoder with SAI while freezing the weights of pre-trained CLM and SC-Encoder modules. Hopefully, we expect that each module could acquire different knowledge with specific pre-training objectives.

CSRL training Our CSRL model is trained only using Chinese CSRL annotations and no additional data is introduced during the CSRL training stage. We train our model to minimize the cross-entropy error for a training sample with label y based on the semantic role encoding $\boldsymbol{l}$,

$$p = \text{softmax}(\boldsymbol{l}_t) \quad \mathcal{L}_{CSRL} = -\sum_{l=1}^{L} y \log p \tag{5}$$

4 Experiments

We evaluate our method from two aspects: 1) the performance of cross-lingual CSRL parser; 2) the usefulness of CSRL parser on conversation-based tasks in target languages. In this section, we describe the data and training details, and provide detailed evaluation results and further discussions.

4.1 Datasets

CSRL data We use DuConv training and development sets with CSRL annotations(Xu et al., 2021) for model training and selection, and use DuConv test set for language in-domain evaluation. Furthermore, we manually collect two CSRL testing datasets[3] for cross-lingual evaluation based on Persona-Chat(Zhang et al., 2018) and CMU-DoG(Zhou et al., 2018), both of which are English conversation datasets. We only explore cross-lingual CSRL on Chinese→English (Zh→En) in this work, and we leave other languages for future work.

[3] More details are described in Appendix B.

Method	DuConv			Persona-Chat			CMU-DoG		
	$\mathrm{F1}_{all}$	$\mathrm{F1}_{cross}$	$\mathrm{F1}_{intra}$	$\mathrm{F1}_{all}$	$\mathrm{F1}_{cross}$	$\mathrm{F1}_{intra}$	$\mathrm{F1}_{all}$	$\mathrm{F1}_{cross}$	$\mathrm{F1}_{intra}$
SimpleBERT	86.54	81.62	87.02	-	-	-	-	-	-
CSRL-BERT	88.46	81.94	89.46	-	-	-	-	-	-
SimpleXLMR	84.89	36.36	84.93	62.96	14.29	63.03	50.54	14.29	58.50
CSRL-XLMR	88.03	78.12	89.33	63.18	18.71	65.05	53.84	34.20	59.78
Back-translation	-	-	-	63.49	13.90	66.67	47.91	27.44	50.92
Fine-tune all parameters									
Ours$_{\text{mBERT}}$	87.20	81.14	88.11	58.38	9.39	61.77	48.13	20.92	52.91
Ours$_{\text{XLM-R}}$	88.35	83.39	89.21	**67.29**	24.29	**70.61**	**61.74**	**60.32**	**62.67**
Ours$_{\text{w/ pretrain}}$	**88.60**	**84.10**	**89.24**	67.23	**25.43**	69.89	59.24	58.94	60.89
Freeze parameters of the language model									
Ours$_{\text{mBERT}}$	87.08	81.46	87.98	59.04	11.23	62.13	48.87	21.78	53.54
Ours$_{\text{XLM-R}}$	88.30	83.38	89.17	65.57	24.11	68.51	**59.60**	56.16	**60.78**
Ours$_{\text{w/ pretrain}}$	**88.60**	**83.72**	**89.27**	**66.75**	**24.13**	**69.44**	58.45	**58.92**	58.82
Ablation study of pre-training objectives									
All objectives	88.60	83.72	89.27	66.75	24.13	69.44	58.45	58.92	58.82
w/o TLM and HPSI	88.07	81.90	89.06	65.07	23.91	68.34	58.23	53.15	59.24
w/o SPI and UOR	87.75	81.56	88.81	68.35	22.86	71.29	58.08	47.93	60.22
w/o SAI	88.00	83.16	89.06	64.74	23.33	67.99	59.94	54.68	61.87

Table 1: Evaluation results on the DuConv, Persona-Chat and CMU-DoG datasets.

Pre-training data For TLM and HPSI objectives which requires parallel data to enhance alignment ability, we choose IWSLT'14 English↔Chinese (En↔Zh) translations[4]. For SPI and UOR objectives whose goal is to model high-level conversational features, we select samples from Chinese conversation dataset (i.e., DuConv) and English conversation datasets (i.e., Persona-Chat and CMU-DoG) with equal probability. For SAI, we borrow CoNLL-2012(Pradhan et al., 2012) which contains standard SRL annotations in Arabic, Chinese and English for pre-training.

We stress that by keeping the sampling balance of Chinese and English data for every pre-training objective and sharing all parameters across the languages, our model would capture task-specific but language-agnostic features.

4.2 Experimental Steup

We implement the model in PyTorch(Paszke et al., 2019), and use the pre-trained language model of multilingual BERT (mBERT) or XLM-RoBERTa (XLM-R) made available by the Transformer library (Wolf et al., 2020) as the backbone. We train the model using AdamW(Loshchilov & Hutter, 2018) with a linear learning rate schedule. For each model, we run five different random seeds and report the average score. More details and hyper-parameters are listed in Table 6 (in Appendix C).

Following previous work(Xu et al., 2021), we evaluate our system on micro-average $\mathrm{F1}_{all}$, $\mathrm{F1}_{cross}$ and $\mathrm{F1}_{intra}$ over the (predicate, argument, label) tuples, wherein we calculate $\mathrm{F1}_{cross}$ and $\mathrm{F1}_{intra}$ over the arguments in the different, or same turn as the predicate. We refer these two types of arguments as *cross*-arguments and *intra*-arguments. For language in-domain evaluation, we compare to *SimpleBERT* (Shi & Lin, 2019) that uses the Chinese BERT as their backbone and simply concatenates the entire dialogue context with the predicate, and *CSRL-BERT* (Xu et al., 2020) that also uses the Chinese BERT as the backbone but attempts to encode the conversation structural information by integrating the dialogue turn and speaker embeddings in the input embedding layer. For cross-lingual evaluation, we compare to *SimpleXLMR* and *CSRL-XLMR* by simply replacing SimpleBERT and CSRL-BERT's backbones with XLM-R. We also compare to a back-translation baseline. Specifically, the test data in English is translated and projected to Chinese annotations using Google Translate (Wu et al., 2016) and the state-of-the-art word alignment toolkit Awesome-align(Dou & Neubig, 2021). Then we feed the translated samples into the pre-trained CSRL model to obtain back-translation results.

[4] https://wit3.fbk.eu/

4.3 Main Results

Table 1 summarizes the results of all compared methods on the DuConv, Persona-Chat and CMU-DoG datasets. We can see that our method outperforms all the baselines by large margins no matter fine-tuning or freezing the language model during the CSRL training stage. First, in contrast to the performance drops of SimpleXLMR and CSRL-XLMR against SimpleBERT and CSRL-BERT on the DuConv dataset, our model using XLM-R as backbone achieves competitive performance to the state-of-the-art CSRL-BERT model across all metrics, especially in terms of F1_{cross} where at least 1.78% gains are obtained. Similar results can also be found on Persona-Chat and CMU-DoG datasets where our cross-lingual model improves all baselines by at least 5.42% on F1_{cross} and 1.00% on F1_{intra}. We think this observation is expected because (1) our model is language-agnostic which makes the cross-lingual transfer easier; (2) our model captures more high-level conversational features in SC-Encoder, thus enhancing the capacities of the model to recognize cross-arguments; (3) rich semantic features are modeled by PA-Encoder, which would improve the capacities of the model to recognize intra-arguments.

Secondly, although our model has achieved outstanding performance on all datasets, further improvements can be observed after incorporating our well-designed pre-training objectives, especially when we freeze the parameters of the language model. However, we find that the performance on the CMU-DoG dataset heavily drops after introducing our pre-training objectives, especially in terms of F1_{intra}. We think this is because the semantic argument spans in CoNLL-2012 are relatively different from those in CMU-DoG, thus leading to the vague boundary detection and performance drop. To verify this assumption, we conduct ablation study by removing SAI from the pre-training stage. Interestingly, we observe substantial improvements over F1_{all} and F1_{intra}, suggesting that pre-training on CoNLL-2012 does hurt the performance on CMU-DoG dataset. Additionally, we find that fine-tuning all parameters leads to slightly better performance than freezing the language model during the CSRL training stage. This finding is also consistent with the previous work (Conia et al., 2021).

Finally, by analyzing the results of ablation experiments, we draw several conclusions: (1) removing TLM and HPSI objectives hurt performance consistently but slightly across all metrics on all datasets; (2) SPI and UOR objectives significantly affect the values of F1_{cross}, especially on two language out-of-domain datasets; (3) SAI objective helps to find intra-arguments on DuConv and Persona-Chat datasets, but might hurt the F1_{intra} performance on CMU-DoG.

4.4 Low-resource cross-lingual CSRL

We evaluate the robustness of our proposed method in low-resource scenario by artificially reducing the size of training set. Specifically, we examine on 10%, 30%, 50% and 70% of training data, respectively. Figure 2 illustrates the F1_{all} scores of these low-resource experiments over all datasets[5]. We can find that our method with pre-training objectives can reach competitive performance just with 30% training data while the vanilla model needs around 50% training data. This result is expected since our model could acquire rich knowledge about dialogue encoding and semantic role identification with the well-designed pre-training objectives. Therefore, we believe that our model is robust to low-resource scenarios, especially after introducing pre-training objectives. This observation is very important and sheds more lights to extend CSRL into low-resource languages.

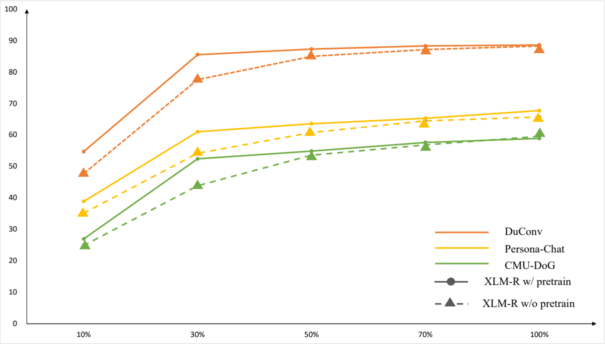

Figure 2: F1_{all} scores of low-resource experiments.

[5] More detailed scores are listed in Appendix D.

4.5 Applications

Xu et al. (2021) has confirmed the usefulness of CSRL by applying CSRL parsing results to two Chinese dialogue tasks, including dialogue context rewriting and dialogue response generation. In the same vein, we also explores whether CSRL could benefit to the same English dialogue tasks.

Question-in-context Rewriting Conversational Question Answering (CQA) (Reddy et al., 2019) is a challenging task that asks multiple questions in an information-seeking dialogue. The current popular approaches to CQA is to model the interactions among the question, the conversation context and reference documents with attention mechanism and then find the answers. However, some questions are very ambiguous and less informative while ellipsis or anaphora occurs in the questions, thus making the model pay vague attentions to text components. For example, in Table 2, the question “who did they play in the playoffs?” cannot be understood without knowing “they” refer to, which can be resolved with the given context. To tackle this problem, Elgohary et al. (2019) proposed a task, named *question-in-context rewriting*, which required the model to resolve the conversational dependencies between the question and the context, and then rewrite the original question into independent one. To this purpose, they collected a dataset **CANARD** by rewriting QuAC questions (Choi et al., 2018) into context-independent paraphrases.

U1	how many games did the colts win?
U2	the Colts$_{\textbf{ARG0}}$ finished with a 12-2 record.
Question	who did they play$_{\textbf{predicate}}$ in the playoffs?
Question$'$	who did the Colts play in the playoffs?

Table 2: One example of question-in-context rewriting.

Since CSRL could identify the predicate-argument structures from the entire conversation, we believe that it can be used to pick up dropped or referred components, and mark important words that semantically related to the question. For example, in Table 2, our CSRL parser can find that the ARG0 of the predicate “play” is “the Colts”. Motivated by this observation, we attempt to borrow CSRL to this task by first recognizing predicate-argument pairs from the conversation context and then encoding them into the rewriter models (Su et al., 2019; Xu et al., 2020).

We adopt the model proposed in (Xu et al., 2020) which directly concatenates the predicate-argument structures, the conversation context and the question as a sequence, and then feeds them into the model with special attention masks. During decoding, the model takes CSRL pairs and the context to generate the rewritten question word by word. The input representation, attention strategies and loss function of our model are same as Xu et al. (2020)’s. We use CANARD dataset to evaluate our method. We initialize the model using base BERT model and use AdamW with a linear learning rate schedule to update parameters. More hyper-parameters are listed in Table 7 (see in Appendix C). We employ the pre-trained cross-lingual CSRL parser to extract predicate-argument pairs from conversations.

Method	BLEU-1	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Seq2Seq	-	-	49.67	-	-	-
Human evaluation	-	-	59.92	-	-	-
Ours$_{\text{wo/ CSRL}}$	69.24	62.93	52.78	65.55	49.72	65.73
Ours$_{\text{w/ CSRL}}$	**70.26**	**64.19**	**54.23**	**67.17**	**51.36**	**67.10**

Table 3: Evaluation results on the dataset of CANARD (Elgohary et al., 2019).

Following previous work, we report the BLEU and ROUGE scores. Table 3 lists the results of our model on CANARD. We can see that even without CSRL information, our implementation could already significantly outperform the baseline method (Bahdanau et al., 2014) over all metrics. After introducing the predicate-argument structures, the performance is further improved by 1.45 BLEU-4, 1.64 ROUGE-2 and 1.37 ROUGE-L. To figure out the reasons of such improvements, we also investigate which type of questions could benefit from CSRL information. By comparing the rewritten questions of different methods, we find that the questions that required information completion, especially those containing referred components, benefit from CSRL most. This observation is naturally in line with our expectation that our CSRL parser could consistently offer essential guidance by recovering dropped or referred text components.

Multi-turn Dialogue Response Generation In addition to rewriting tasks that are heavily affected by omitted components, we also explore the usefulness of CSRL to *multi-turn dialogue response generation*, one of the main challenges in dialogue community. In contrast to the single-turn dialogue response generation, multi-turn dialogues suffers more frequently occurred ellipsis and anaphora which would lead to vague context representations. However, previous approaches to tackle this problem is to simply concatenate the multi-turn dialogues into a "long" sequence and adopt a hierarchical structure to implicitly model the relations among words and utterances. Thanks to the nature that CSRL can extract semantic pairs from the entire conversation, we can highlight the words pick up by the CSRL parser, and then teach the model to pay more attention on those words which would hold more semantic information.

Our model for response generation is analogous to Dong et al. (2019); Xu et al. (2020) which can flexibly support both bi-directional encoding and uni-directional decoding via special self-attention masks. Specifically, we first employ the pre-trained cross-lingual CSRL parser to analyze the last utterance, and then we concatenate the extracted predicate-argument pairs with the context and target response into a sequence. We feed the sequence into our model for training; during decoding, our model takes semantic information and the context as input to generate the response word by word. The input representation, attention strategies and loss function are same as the rewriter model's.

We conduct evaluations on **Persona-Chat** (Zhang et al., 2018), an English persona-based dialogue dataset containing 162,064 utterances over 10,907 dialogues. Since our goal is to verify the effectiveness of CSRL to multi-turn dialogue response generation, we drop the persona knowledge in our experiments and directly compare the performance after introducing CSRL information. Analogous to rewriting experiments, we initialize the mode using base BERT model and use AdamW with a linear learning rate schedule to update the parameters. More hyper-parameters are listed in Table 8.

Following previous work, we report BLEU-1/2 and Distinct-1/2 scores. Table 4 summarize the results of multi-turn dialogue response generation on Persona-Chat dataset. We can see that our implementation significantly outperforms the baseline method (Bahdanau et al., 2014) even without CSRL information. After introducing CSRL information, we obtain further gains across all metrics. Apart from automatic evaluation criteria, we also conduct human evaluation. Specially, we randomly select 200 generated responses for each method, and then recruit three annotators to evaluating the coherence and informativeness of the response against the conversation context by giving a score ranging from 1(worst) to 5(best). We find that our model with CSRL wins in 35% cases, and ties with the vanilla model in around 55% cases. With more careful comparisons, we find that the responses that contains entities mentioned in histories benefit from CSRL information most. We think this is because none-phrases are more likely to be recognized as semantic arguments by CSRL parser, and then receive more attentions during encoding.

Method	B1/2	D1/2	Human
Seq2Seq	0.138/0.069	0.051/0.094	2.72
Ours$_{\text{wo/ CSRL}}$	0.188/0.113	0.114/0.217	3.02
Ours$_{\text{w/ CSRL}}$	**0.195/0.122**	**0.116/0.223**	**3.16**

Table 4: Evaluation results on Persona-Chat.

According to the impressive experimental results of question-in-context rewriting and multi-turn dialogue response generation, we firmly believe that CSRL information is helpful to English downstream dialogue tasks. In addition, our cross-lingual CSRL parser is also capable to analyze English conversations and generate reasonable predicate-argument structures.

5 Conclusion and Future Work

In this work, we propose a simple but effective model with five well-designed pre-training objectives to perform zero-shot cross-lingual CSRL. Experimental results show that our model achieves outstanding performance on all test sets. Further explorations on low-resource scenario also demonstrate the robustness of our method. In addition, we also confirm the effectiveness of CSRL to English dialogue tasks by introducing CSRL information into these tasks. Future work can be conducted to further improve cross-lingual CSRL performance or to explore more applications of CSRL.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pp. 152–164, 2005.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, 2018.

Simone Conia and Roberto Navigli. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1396–1410, 2020.

Simone Conia, Andrea Bacciu, and Roberto Navigli. Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 338–351, 2021.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020.

Angel Daza and Anette Frank. Translate and label! an encoder-decoder approach for cross-lingual semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 603–615, 2019.

Angel Daza and Anette Frank. X-srl: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3904–3914, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 13063–13075, 2019.

Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2112–2128, 2021.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5918–5924, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1605. URL `https://aclanthology.org/D19-1605`.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22:1–48, 2021.

Hao Fei, Meishan Zhang, and Donghong Ji. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7014–7026, 2020a.

Hao Fei, Meishan Zhang, Fei Li, and Donghong Ji. Cross-lingual semantic role labeling with model transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2427–2437, 2020b.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3682–3692, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.285.

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

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499, 2020.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2649–2663, 2020.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1297–1303, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3836–3845, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1373.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037, 2019.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pp. 1–40, 2012.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6924–6931, 2019.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, Mao-song Sun, et al. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327, 2018.

Tom Sherborne and Mirella Lapata. Zero-shot cross-lingual semantic parsing. *arXiv preprint arXiv:2104.07554*, 2021.

Tom Sherborne, Yumo Xu, and Mirella Lapata. Bootstrapping a crosslingual semantic parser. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 499–517, 2020.

Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 22–31, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1003. URL `https://aclanthology.org/P19-1003`.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. On learning universal representations across languages. In *International Conference on Learning Representations*, 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *EMNLP (Demos)*, 2020.

Han Wu, Kun Xu, Linfeng Song, Lifeng Jin, Haisong Zhang, and Linqi Song. Domain-adaptive pretraining methods for dialogue understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 665–669, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.84.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3794–3804, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1369. URL `https://aclanthology.org/P19-1369`.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Kun Xu, Haochen Tan, Linfeng Song, Han Wu, Haisong Zhang, Linqi Song, and Dong Yu. Semantic role labeling guided multi-turn dialogue rewriter. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6632–6639, 2020.

Kun Xu, Han Wu, Linfeng Song, Haisong Zhang, Linqi Song, and Dong Yu. Conversational semantic role labeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Yi Xu and Hai Zhao. Dialogue-oriented pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2663–2673, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.235.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

Zhuosheng Zhang and Hai Zhao. Structural pre-training for dialogue comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5134–5145, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.399.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 708–713, 2018.

A Hard Parallel Sentence Identification Sampling

Following previous work (Robinson et al., 2020; Wei et al., 2020) which suggests that contrastive learning of representations benefits from hard negative samples, we also try to select hard negative samples for PSI task based on n-gram similarity and text perturbation. Specifically, for each sentence, we calculate its n-gram similarity scores to other sentences, where $n = 1, 2, 3, 4$, and then we select the sentence with the highest score at each gram as the candidate sentence; additionally, we construct the corrupted sentence as the candidate by token deletion, token replacement and token order permutation. Finally, we sample from the candidate set created by n-gram similarity at 40% time and from the candidate set created by text perturbation at 60% time.

B Dataset Statistics

Dataset	language	#dialogue	#utterance	#predicate	#tokens per utterance	cross ratio
DuConv	ZH	3,000	27,198	33,673	10.56	21.89%
Persona-Chat	EN	50	2,669	477	17.96	17.74%
CMU-DoG	EN	50	3,217	450	12.57	7.41%

Table 5: Statistics of the annotations on DuConv, NewsDialog and PersonalDialog.

Following the instructions in Xu et al. (2021), we manually collect two out-of-domain CSRL test sets based on English dialogue datasets Persona-Chat (Zhang et al., 2018) and CMU-DoG (Zhou et al., 2018). The statistics of the datasets are listed in Table 5.

C Hyper-parameters

We list the hyper-parameters of CSRL experiments (Table 6), rewriting experiments (Table 7) and response experiments (Table 8) below.

D Low-resource experiments

We list the detailed results of low-resource experiments at Table 9.

Name	Value
Language model	xlm-roberta-base
Hidden state size	512
Word-level encoder layers	2
Pred.-arg encoder layers	1
Batch size per GPU	24
Max learning rate	5e-5
Min learning rate	1e-5
Max *lr* for LM fine-tuning	1e-5
Min *lr* for Lm fine-tuning	1e-6
Max sequence length	512
Max training epochs	50
Max training steps	15000
Early-stop patience	10

Table 6: Hyper-parameters in CSRL experiments.

Name	Value
Language model	bert-base-cased
Hidden state size	768
Batch size per GPU	16
Max learning rate	3e-5
Min learning rate	1e-5
Max sequence length	512
Max decode length	32
Max training epochs	20
Early-stop patience	5

Table 7: Hyper-parameters in rewriting experiments.

Name	Value
Language model	bert-base-cased
Hidden state size	768
Batch size per GPU	16
Max learning rate	5e-5
Min learning rate	3e-5
Max sequence length	512
Max decode length	64
Max training epochs	20
Early-stop patience	5

Table 8: Hyper-parameters in response generation experiments.

Method	DuConv			Persona-Chat			CMU-DoG		
	$\mathrm{F1}_{all}$	$\mathrm{F1}_{cross}$	$\mathrm{F1}_{intra}$	$\mathrm{F1}_{all}$	$\mathrm{F1}_{cross}$	$\mathrm{F1}_{intra}$	$\mathrm{F1}_{all}$	$\mathrm{F1}_{cross}$	$\mathrm{F1}_{intra}$
$\text{Ours}_{\text{XLM-R / 10\% data}}$	47.73	45.60	47.90	35.14	6.51	36.97	24.88	22.58	25.31
$\text{Ours}_{\text{XLM-R / 30\% data}}$	77.62	72.00	78.81	54.20	16.19	56.91	43.88	42.26	44.86
$\text{Ours}_{\text{XLM-R / 50\% data}}$	85.03	78.84	86.34	60.78	22.87	63.70	53.57	48.97	55.37
$\text{Ours}_{\text{XLM-R / 70\% data}}$	87.18	81.61	88.20	64.51	23.71	67.43	56.87	53.61	58.25
$\text{Ours}_{\text{XLM-R / pre-train / 10\% data}}$	54.74	56.33	53.70	38.91	8.71	41.08	26.96	24.66	26.84
$\text{Ours}_{\text{XLM-R / pre-train / 30\% data}}$	85.56	79.72	86.57	61.02	18.46	63.50	52.43	52.67	52.88
$\text{Ours}_{\text{XLM-R / pre-train / 50\% data}}$	87.31	82.31	88.07	63.60	25.04	65.94	54.87	50.82	56.20
$\text{Ours}_{\text{XLM-R / pre-train / 70\% data}}$	88.31	83.07	89.08	65.32	22.12	68.02	57.64	56.32	58.26

Table 9: Low-resource experiments on the DuConv, Persona-Chat and CMU-DoG datasets.