Zero-shot Cross-lingual Conversational Semantic Role Labeling

Anonymous ACL submission

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. Experimental results show that our cross-lingual model not only outperforms baselines by large margins but it is also robust to low-resource scenarios. More importantly, we confirm the usefulness of CSRL to English conversational tasks such as question-in-context rewriting and multi-turn dialogue response generation by incorporating the CSRL information into the downstream conversation-based models. We believe this finding is significant and will facilitate the research of English dialogue tasks which suffer the problems of ellipsis and anaphora.

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 and Frank, 2020) 2) borrowing machine translation and word alignment techniques to transfer the dataset in source language into target language (Daza and Frank, 2019; Fei et al., 2020a) 3) zero-shot transfer learning with multilingual pre-trained language model (Rijhwani et al., 2019; Sherborne and 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 and Lample, 2019; Conneau et al., 2020) on cross-lingual tasks, including machine translation (Lin et al., 2020; Liu et al., 2020b; Fan et al., 2021; Chen et al., 2021), semantic role labeling (SRL) (Conia and Navigli, 2020; Conia et al., 2021) and semantic parsing (Fei et al., 2020b; Sherborne et al., 2020; Sherborne and 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 and Zhao, 2021; Zhang and Zhao, 2021) and semantic arguments identification (Wu et al., 2021a; 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 summarize our contributions as follows: **(1)** We propose a simple but effective model which consists of three modules, namely cross-lingual language model (CLM), structure-aware conversation encoder (SA-Encoder) and predicate-argument encoder (PA-Encoder), and five well-designed pre-training objectives. Our model implicitly learns language-agnostic, conversational structure-aware and semantically rich representations to perform zero-shot cross-lingual CSRL. **(2)** Experiments show that our proposed method outperforms all baselines and achieves impressive cross-lingual performance no matter whether incorporating the pre-training. **(3)** We confirm the usefulness of CSRL to English dialogue tasks including question-in-context rewriting and response generation. We believe this finding is important and will facilitate the research of English dialogue tasks that suffer ellipsis and anaphora. **(4)** We will release our code, the new annotated English CSRL test sets and checkpoints of our best models to facilitate the further research.

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 and 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 natural language generation (Shen et al., 2018) and understanding (Liu et al., 2019; Lauscher et al., 2020; Sherborne and Lapata, 2021). In this work, we claim our method is zero-shot because no non-Chinese CSRL annotations are seen during the CSRL training stage. For decoding, we directly use the cross-lingual CSRL model trained on Chinese CSRL data to analyze conversations in other languages. 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 Chinese 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 languages has not been confirmed yet due to the lack of cross-lingual CSRL models. In this work, we primarily focus on the latter problem and propose a simple but effective model to perform cross-lingual CSRL. We would like to distinct our work from the concurrent work (Wu et al., 2021b) which purely focuses on improving the CSRL performance. Wu et al. (2021b) try to model predicate-aware representations which could benefit to monolingual CSRL task, but hurt the cross-lingual performance, because the relative positions of the predicates may differ from language to language.

3 Methodology

Following Xu et al. (2021), we solve the CSRL task as a sequence labeling problem. Our goal is to find the arguments over the entire dialogue with the given predicate and additional information such as turn and speaker role indicators.

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 $\mathbf{e}$ by concatenating the hidden states of the four top-most layers of the language model.

Structure-aware Conversation Encoder (SC-Encoder) Different from standard SRL(Carreras

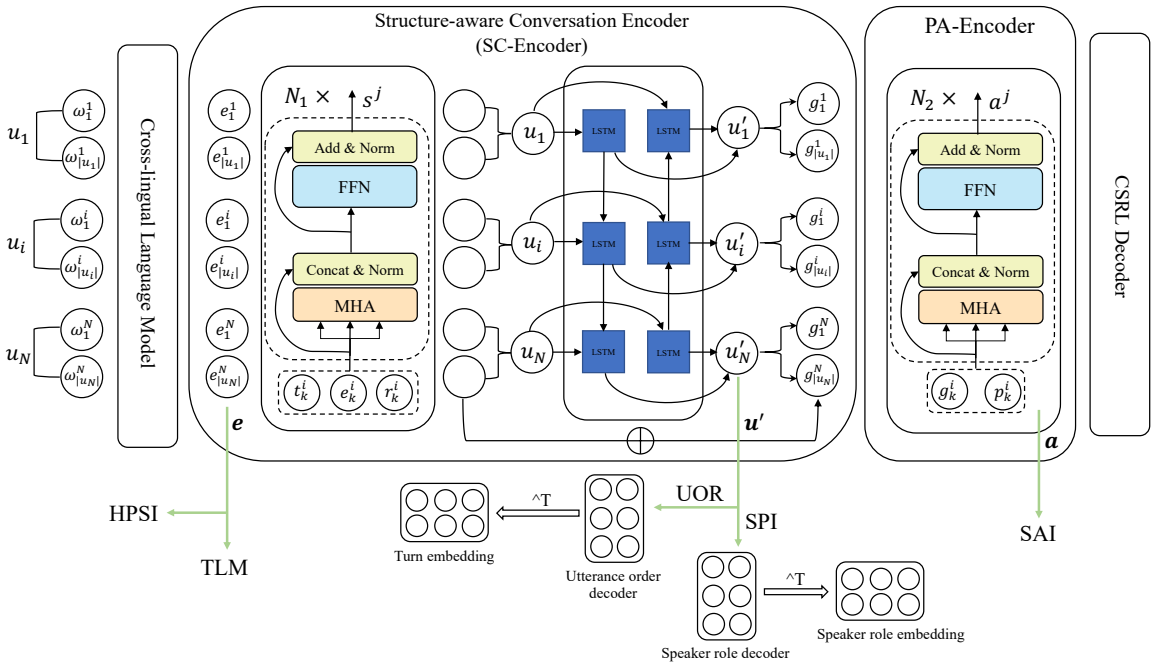

Figure 1: Overall model architecture.

and Màrquez, 2005), CSRL requires the models to find arguments from no only the current turn, but also previous turns, leading to more challenges of dialogue modeling. 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:

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

where $\boldsymbol{s}_{(i,k)}^j$ is the timestep encoding of k-th token in i-th utterance from j-th word-level encoder layer while $j \in (0, \ldots, N_1)$, $\oplus$ represents vector concatenation, and MTRANS is the **M**odified **Trans**former encoder layer. Concretely, we replace the `[Add]` operation in the first residual connection layer 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 resulting utterance representations $\boldsymbol{u}$ through a stack of Bi-LSTM (Hochreiter and 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:

$$\boldsymbol{g}_k^i = \text{Swish}(\mathbf{W}^g[\boldsymbol{s}_{(i,k)}^{N_1} \oplus \boldsymbol{u}_i'] + \mathbf{b}^g) \quad (2)$$

where $\text{Swish}(x) = x \cdot \text{sigmoid}(x)$ is a non-linear activation function, $\boldsymbol{s}_{i,k}^{N_1}$ 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 couple that appears in the conversation. Similar with the word-level encoder, we use a stack of MTRANS layers to implement this encoder. Formally, with denoting predicate embedding as $\boldsymbol{p} = (\boldsymbol{p}_1^1, ..., \boldsymbol{p}_k^i, ..., \boldsymbol{p}_{|u_N|}^N)$, the model calculates the predicate-specific argu-

[1]More details about MTRANS in Appendix B.

ment encodings as follows:

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

where $\boldsymbol{g}^{i}_{k}$ is the token embedding from conversation encoder, $\boldsymbol{p}^{i}_{k}$ is the corresponding predicate indicator embedding, and $\boldsymbol{a}^{j}_{(i,k)}$ is the argument encoding of k-th token in i-th utterance from j-th encoder layer while $j \in (0, \ldots, N_2)$. Finally, we obtain the semantic role encoding $\boldsymbol{l}$ using the resulting argument encodings from the last layer of the predicate-argument encoder:

$$\boldsymbol{l}^{i}_{k} = \text{Swish}(\mathbf{W}^{l}\boldsymbol{a}^{N_2}_{(i,k)} + \mathbf{b}^{l}) \tag{4}$$

In particular, we emphasize that our proposed model is mostly language-agnostic since we do not explicitly introduce any language-specific knowledge such as word order, part-of-speech tags or dependent relations, and only incorporate the predicate indicator that might contain some language-specific information in the semantic module, which would not affect the latent space alignment and dialogue modeling.

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 and 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.

Besides improving word-level alignment ability by TLM, we also attempt 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 and 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 closer the negative sample is to the positive sample, 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 and Zhao, 2021; Gu et al., 2021) and semantic representations (Wu et al., 2021a), we tend to explicitly model speaker dependency and temporal dependency in the conversation, both of which have been proven to be critical to CSRL task (Xu et al., 2021).

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 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. The goal of SPI is to predict the masked speaker roles according to the corrupted speaker indicators and context. Secondly, we borrow utterance order permutation (UOR) 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.

In practice, we drop the dialogue turn embedding here to avoid temporal information leakage. We use the sequentially informed utterance representations $\boldsymbol{u}'$ as the input of speaker role and utterance order decoders, and pre-train SC-Encoder using the combination of SPI and UOR. After the pre-training of this stage, we respectively employ the transposed speaker role and utterance order decoders as the speaker role and dialogue turn embedding matrices during the CSRL training stage.

[2]Details in Appendix A

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 strengthen the correlations between the predicate and its arguments with the help of external standard SRL corpus, i.e., CoNLL-2012. Specifically, for each SRL sample, we only focus on those arguments, including ARG0-4, ARG-LOC, ARG-TMP and ARG-PRP, all of which are defined in both SRL and CSRL tasks. The model is encouraged to find the textual spans of these arguments with the given predicate. We believe 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 because they do not have any conversational characteristics. We directly feed the word-level context representations $\boldsymbol{s}$ into PA-Encoder, and then use the argument encodings $\boldsymbol{a}$ to make classifications.

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 $\boldsymbol{l}$,

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

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.

4.1 Datasets

CSRL data We use the same split as Xu et al. (2021) where DuConv annotations are splitted into 80%/10%/10% as train/dev/in-domain test set. Furthermore, we manually collect two CSRL test sets[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. Note that we only explore cross-lingual CSRL on Chinese→English (Zh→En) here, and we leave other languages for future work.

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 the Chinese and English SRL annotations from CoNLL-2012(Pradhan et al., 2012).

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 Setup

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 and 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 G).

Following previous work(Xu et al., 2021), we evaluate our system on micro-average F1_{all}, F1_{cross} and F1_{intra} over the (predicate, argument, label) tuples, wherein we calculate F1_{cross} and 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 and Lin, 2019), *CSRL-BERT* (Xu et al., 2021) and *CSAGN* (Wu et al., 2021b), all

[3]More details are described in Appendix C.
[4]https://wit3.fbk.eu/

Method	Trainable parameters	DuConv			Persona-Chat			CMU-DoG		
		$F1_{all}$	$F1_{cross}$	$F1_{intra}$	$F1_{all}$	$F1_{cross}$	$F1_{intra}$	$F1_{all}$	$F1_{cross}$	$F1_{intra}$
SimpleBERT	117 M	86.54	81.62	87.02	-	-	-	-	-	-
CSRL-BERT	147 M	88.46	81.94	89.46	-	-	-	-	-	-
CSAGN	163 M	89.47	84.57	90.15	-	-	-	-	-	-
SimpleXLMR	292 M	84.75	63.44	85.12	62.96	14.29	63.03	50.54	14.29	58.50
CSRL-XLMR	320 M	88.03	78.12	89.33	63.18	18.71	65.05	53.84	34.20	59.78
CSAGN-XLMR	338 M	88.52	82.45	89.98	63.02	17.82	64.97	52.73	30.11	58.91
Back-translation	-	-	-	-	63.49	13.90	66.67	47.91	27.44	50.92
Fine-tune all parameters										
$\text{Ours}_{\text{mBERT}}$	272 M	87.20	81.14	88.11	58.38	9.39	61.77	48.13	20.92	52.91
$\text{Ours}_{\text{XLM-R}}$	372 M	88.35	83.39	89.21	**67.29**	24.29	**70.61**	**61.74**	**60.32**	**62.67**
$\text{Ours}_{\text{w/ pretrain}}$	372 M	**88.60**	**84.10**	**89.24**	67.23	**25.43**	69.89	59.24	58.94	60.89
Freeze parameters of the language model										
$\text{Ours}_{\text{mBERT}}$	180 M	87.08	81.46	87.98	59.04	11.23	62.13	48.87	21.78	53.54
$\text{Ours}_{\text{XLM-R}}$	180 M	88.30	83.38	89.17	65.57	24.11	68.51	**59.60**	56.16	**60.78**
$\text{Ours}_{\text{w/ pretrain}}$	180 M	**88.60**	**83.72**	**89.27**	**66.75**	**24.13**	**69.44**	58.45	**58.92**	58.82
Ablation studies										
All objectives	-	88.60	83.72	89.27	66.75	24.13	69.44	58.45	58.92	58.82
w/o TLM & HPSI	-	88.07	81.90	89.06	65.07	23.91	68.34	58.23	53.15	59.24
w/o SPI & 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
w/ end2end pre-training	-	87.28	81.02	88.73	64.37	21.17	67.77	57.86	50.40	58.20
$\text{Ours}_{\text{XLM-R}}$	-	88.30	83.38	89.17	65.57	24.11	68.51	59.60	56.16	60.78
w/o SC-Encoder	-	88.02	79.11	89.05	63.12	17.55	66.70	57.72	50.42	58.03
w/o PA-Encoder	-	88.10	81.32	88.78	64.05	22.38	64.82	58.24	54.00	59.23
w/o MTRANS	-	88.25	83.01	89.08	65.27	23.10	68.38	58.58	55.41	59.98

Table 1: Evaluation results on the DuConv, Persona-Chat and CMU-DoG datasets. Scores in GRAY are from the concurrent work (Wu et al., 2021b).

of which employ the Chinese pre-trained language model as the backbone. For cross-lingual evaluation, we compare to *SimpleXLMR*, *CSRL-XLMR* and *CSAGN-XLMR* by simply replacing the BERT backbones of those models with XLM-R. Additionally, 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 and Neubig, 2021). We feed the translated samples into CSAGN to obtain the back-translation results.

4.3 Main Results

Table 1 summarized the results of all compared methods on DuConv, Persona-Chat and CMU-DoG datasets.

Firstly, we can see that our method achieves competitive performance over all datasets, especially in cross-lingual scenario where our method outperforms the baselines by a large margin no matter fine-tuning or freezing the language model during the CSRL training stage. Although CSAGN exceeds our method on DuConv test set, it fails to work well in cross-lingual scenario. We think the reasons are (1) it heavily relies on rich features of the Chinese pre-trained language model (2) it is overfitting on the predicate-aware information. Superior to CSAGN, our model with the multilingual backbone achieves outstanding performance on both language in-domain and cross-lingual datasets. 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 good performance over all datasets, further improvements can be observed after incorporating our well-designed pre-training objectives, especially when freezing the parameters of the language model. Exceptionally, we find that the performance on the CMU-DoG dataset heavily drops after introducing the pre-training objectives, especially in terms

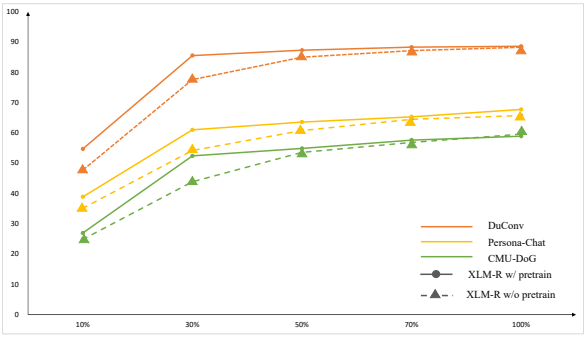

Figure 2: $F1_{all}$ scores of low-resource experiments on DuConv, Persona-Chat and CMU-DoG.

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. Furthermore, we also find that fine-tuning all parameters leads to slightly better performance than freezing the language model during the CSRL training stage. This finding is consistent with the previous work (Conia et al., 2021). However, we do not think this improvement is efficient since it consumes much more computation resources. To this end, we are more focused on the performance using the frozen language model.

Finally, by analyzing the results of ablation studies, we draw several conclusions: (1) removing PA-Encoder or MTRANS or TLM & HPSI objectives hurt performance consistently but slightly; (2) SPI, UOR objectives and SC-Encoder significantly affect the values of $F1_{cross}$, especially on two cross-lingual datasets; (3) SAI objective helps to find intra-arguments on DuConv and Persona-Chat, but might hurt the $F1_{intra}$ score on CMU-DoG; (4) hierarchical pre-training is superior to end-to-end pre-training which simultaneously optimizes all auxiliary objectives.

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 (Detail scores in Table 10). We can find that our method

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

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

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 method is robust to low-resource scenarios, especially after introducing pre-training objectives. This observation sheds more lights to extend CSRL into low-resource languages.

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 explore whether CSRL could benefit to the same English dialogue tasks.

Question-in-context Rewriting *Question-in-context rewriting* (Elgohary et al., 2019) is a challenging task which requires the model to resolve the conversational dependencies between the question and the context, and then rewrite the original question into independent one. This is an example in Table 2. The question "who did they play in the playoffs?" cannot be independently understood without knowing "they" refer to, but it can be resolved with the given context.

Since the CSRL models can identify the predicate-argument structures from the entire conversation, we believe that it can help this rewriting task by searching the dropped or referred components from the context. 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 help the question rewriting with the context. We first employ the pre-trained cross-lingual CSRL parser to extract predicate-argument pairs from conversations. Then, we adopt the model proposed in (Xu et al., 2020) to achieve the rewriting. More details about the model are in Appendix E.

We evaluate on CANARD (Elgohary et al., 2019) which is a widely used English question rewriting dataset, and report the BLEU scores. Table 3 lists

Method	B1	B2	B4
Seq2Seq	-	-	49.67
SARG(Huang et al., 2020)	-	-	**54.80**
RUN(Liu et al., 2020a)	**70.50**	61.20	49.10
Human evaluation	-	-	59.92
$\text{Ours}_{\text{wo/ CSRL}}$	69.24	62.93	52.78
$\text{Ours}_{\text{w/ CSRL}}$	70.26	**64.19**	54.23

Table 3: Evaluation results on the dataset of CANARD.

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

Table 4: Evaluation results on Persona-Chat.

the results of our model on CANARD. We can see that our implementation achieves competitive performance against the state-of-the-art rewriting models, i.e., SARG (Huang et al., 2020) and RUN (Liu et al., 2020a), and significantly outperforms the baseline method (Bahdanau et al., 2014). However, in this part, we are more focused on the improvements after introducing CSRL information. We find that the scores across all metrics are improved with the aid of CSRL. To figure out the reasons of these improvements, we investigate which type of questions could benefit from CSRL information. By comparing the rewritten questions of different methods, we find that the questions that requires information completion, especially those containing referred components (around 15% cases), 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 the rewriting task that is 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 single-turn dialogue response generation, multi-turn dialogues suffers more frequently occurred ellipsis and anaphora, which leads to vague context representations. To this end, we attempt to employ CSRL to build better context representations. Specifically, we highlight the words picked up by the CSRL parser, and then teach the model to pay more attention on those words which would hold more semantic information. We first employ the pre-trained cross-lingual CSRL parser to analyze the latest utterance, and then concatenate the extracted predicate-argument pairs with the context and target response into a sequence. Our model for response generation is borrowed from Dong et al. (2019) which can flexibly support both bi-directional encoding and unidirectional decoding via special attention masks.

We evaluate on **Persona-Chat** (Zhang et al., 2018) which is an English persona-based dialogue dataset containing 162,064 utterances over 10,907 dialogues, and report BLEU-1/2 and Distinct-1/2 scores. Note that our goal is to verify the effectiveness of CSRL to multi-turn dialogue response generation, so we drop the persona knowledge in our experiments and directly compare the performance with and without CSRL information. Table 4 summarize the results of 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. Specifically, we randomly select 200 generated responses for each method, and then recruit three annotators to evaluate 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 analysis, 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.

With the impressive experimental results on these two tasks, we firmly believe that CSRL information is helpful to English downstream dialogue tasks. In addition, our cross-lingual CSRL parser is also proven to be capable to analyze English conversations and generate reasonable predicate-argument structures.

5 Conclusion

In this work, we propose a simple but effective model with five pre-training objectives to perform zero-shot cross-lingual CSRL, and 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 explore more applications of CSRL.

References

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

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

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. 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*, pages 338–351.

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

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

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

Angel Daza and Anette Frank. 2019. 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)*, pages 603–615.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

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

Zi-Yi Dou and Graham Neubig. 2021. 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*, pages 2112–2128.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. 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)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

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

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

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

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. 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)*, pages 3682–3692, Online. Association for Computational Linguistics.

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

Mengzuo Huang, Feng Li, Wuhe Zou, Weidong Zhang, and Weidong Zhang. 2020. Sarg: A novel semi autoregressive generator for multi-turn incomplete utterance restoration. *arXiv preprint arXiv:2008.01474*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. 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)*, pages 4483–4499.

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

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020a. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2846–2857.

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

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. 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)*, pages 1297–1303.

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

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy. Association for Computational Linguistics.

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

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

Han Wu, Kun Xu, Linfeng Song, Lifeng Jin, Haisong Zhang, and Linqi Song. 2021a. 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)*, pages 665–669, Online. Association for Computational Linguistics.

Han Wu, Kun Xu, and Linqi Song. 2021b. CSAGN: Conversational structure aware graph network for conversational semantic role labeling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2312–2317, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

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

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

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

Yi Xu and Hai Zhao. 2021. Dialogue-oriented pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2663–2673, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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)*, pages 2204–2213.

Zhuosheng Zhang and Hai Zhao. 2021. 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)*, pages 5134–5145, Online. Association for Computational Linguistics.

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

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 Modified Transformer Encoder Layer

To overcome the information forgetting of hierarchical models, we attempt to modify the standard Transformer to better reserve the information from the previous layers. In specific, we try following variants:

- **MTRANS**. Replacing the `[Add]` operation in the **first** residual connection layer with `[Concat]`.
- **LATER-MTRANS**. Replacing the `[Add]` operation in the **second** residual connection layer with `[Concat]`.
- **BOTH-MTRANS**. Replacing the `[Add]` operations in **both** the first and second residual connection layers with `[Concat]`.

Our intuition of substituting the summation with concatenation is that the residual layer with concatenation would introduce additional parameters, and we expect these additional parameters to retain more history information. As shown in Table 1, we obtain some gains while using MTRANS. Additionally, we also report the $F1_{all}$ scores on DuConv/Persona-Chat/CMU-DoG datasets while using LATER-MTRANS and BOTH-MTRANS here. LATER-MTRANS achieves 88.18/65.32/58.44 points, and BOTH-MTRANS achieves 88.40/66.12/59.72 points against the standard Transformer achieving 88.25/65.27/58.58 points. Although BOTH-MTRANS achieves the best performance, we finally choose MTRANS since BOTH-MTRANS brings a large volume of additional parameters which leads to a huge model size while the increasing of model parameters caused by MTRANS is acceptable.

C Dataset Statistics

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). Specifically, we also annotate the arguments ARG0-4, ARG-TMP, ARG-LOC and ARG-PRP and require that the labeled arguments can only appear in the current turn or the previous turns. We employ three annotators who have studied Chinese CSRL annotations for a period time before this annotation. The first two annotators are required to label all cases and any disagreement between them is solved by the third annotator. The statistics of the datasets are listed in Table 5.

D Baselines

We compare to following baseline models,

1. **SimpleBERT/SimpleXLMR** (Shi and Lin, 2019). It uses the Chinese BERT or XLM-R as the backbone and simply concatenates the entire dialogue context with the predicate.
2. **CSRL- BERT/XLMR** (Xu et al., 2021). It uses the Chinese BERT or XLM-R as the backbone but attempts to encode the conversation structural information by integrating the dialogue turn and speaker embeddings in the input embedding layer.
3. **CSAGN/CSAGN-XLMR** (Wu et al., 2021b). It uses the Chinese BERT or XLM-R as the backbone and employ the relational graph neural network to model predicate- and speaker-aware dependencies. We implement this baseline based on the code `https://github.com/hahahawu/CSAGN`.

E Application Models

Rewriting Model. 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

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.

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 initialize the model using the base BERT model and use AdamW with a linear learning rate schedule to update parameters. We list the hyper-parameters in Table 7.

Response Generation Model. Our model for response generation is analogous to Dong et al. (2019) which can flexibly support both bi-directional encoding and uni-directional decoding via special self-attention masks. Specifically, we concatenate the extracted predicate-argument pairs with the context and target response into a sequence, and then feed the sequence into the encoder 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 initialize the model using the base BERT model and use AdamW with a linear learning rate schedule to update parameters. We list the hyper-parameters in Table 8.

F More Experimental Results

We report some more detailed experimental results here. Table 9 summarize the standard deviations of the main evaluation results on three datasets. Table 10 gives the detailed scores of low-source experiments.

G Hyper-parameters

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

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		
	$F1_{all}$	$F1_{cross}$	$F1_{intra}$	$F1_{all}$	$F1_{cross}$	$F1_{intra}$	$F1_{all}$	$F1_{cross}$	$F1_{intra}$
SimpleXLMR	± 0.32	± 0.61	± 0.20	± 1.16	± 1.95	± 0.72	± 0.68	± 1.73	± 0.17
CSRL-XLMR	± 0.25	± 0.63	± 0.13	± 1.24	± 1.40	± 0.87	± 0.52	± 0.99	± 0.40
CSAGN-XLMR	± 0.27	± 0.32	± 0.21	± 1.31	± 2.18	± 0.78	± 0.54	± 1.04	± 0.43
Back-translation	-	-	-	± 0.67	± 1.12	± 0.56	± 0.42	± 0.55	± 0.44
Fine-tune all parameters									
$\text{Ours}_{\text{mBERT}}$	± 0.31	± 0.46	± 0.24	± 0.94	± 1.23	± 0.70	± 0.51	± 1.12	± 0.32
$\text{Ours}_{\text{XLM-R}}$	± 0.16	± 0.21	± 0.13	± 0.71	± 0.82	± 0.49	± 0.33	± 0.47	± 0.26
$\text{Ours}_{\text{w/ pretrain}}$	± 0.13	± 0.19	± 0.12	± 0.65	± 0.79	± 0.45	± 0.74	± 1.10	± 0.72
Freeze parameters of the language model									
$\text{Ours}_{\text{mBERT}}$	± 0.41	± 0.64	± 0.34	± 1.62	± 2.23	± 1.32	± 1.15	± 1.20	± 1.22
$\text{Ours}_{\text{XLM-R}}$	± 0.23	± 0.38	± 0.17	± 1.07	± 1.41	± 1.10	± 0.82	± 1.30	± 0.87
$\text{Ours}_{\text{w/ pretrain}}$	± 0.23	± 0.31	± 0.18	± 1.00	± 1.25	± 0.90	± 1.12	± 1.35	± 1.20

Table 9: The standard deviations of the main evaluation results on the DuConv, Persona-Chat and CMU-DoG datasets. For SimpleBERT, CSRL-BERT and CSAGN, we directly copy their evaluation scores from (Wu et al., 2021b), so we do not report the standard deviations here.

Method	DuConv			Persona-Chat			CMU-DoG		
	$F1_{all}$	$F1_{cross}$	$F1_{intra}$	$F1_{all}$	$F1_{cross}$	$F1_{intra}$	$F1_{all}$	$F1_{cross}$	$F1_{intra}$
$\text{Ours}_{\text{XLM-R / 10\% data}}$	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 10: Low-resource experiments on the DuConv, Persona-Chat and CMU-DoG datasets.