Parallel Corpora Alignment Framework for Multilingual and Robust Automatic Dialogue Evaluation

Xinglin Wang ♠ **Jiayi Shi** ♠ **Peiwen Yuan** ♠ **Kan Li** ♡
School of Computer Science & Technology, Beijing Institute of Technology
{wangxinglin, shijiayi, yuanpw, likan}@bit.edu.cn

Abstract

Open-domain automatic dialogue evaluation plays an important role in dialogue systems. While recent efforts are being put into making learning-based evaluation metrics correlate better with human evaluation, robust metrics for parallel corpora and multiple domains remain unexplored. Parallel corpora refer to corpora that express the same idea in different ways (e.g., translation, paraphrasing and back-translation). In this paper, we propose **P**arallel **C**orpora **A**lignment **F**ramework (PCAF), which improves the consistency and robustness of model evaluation on parallel corpora. Firstly, parallel corpora are aligned in semantic space through parallel-corpora-aligned contrastive learning. Then, parallel-corpora-aligned distillation on multiple datasets is applied to further improve model's generalization ability across multiple data domains. Our approach ranks second on the final test data of DSTC11 track4 sub-task1 ("Multilingual Automatic Evaluation Metrics", turn-level) and third on the sub-task2 ("Robust Automatic Evaluation Metrics", turn-level), which proves the strong generalization ability and robustness of our proposed approach.

1 Introduction

Open-domain automatic dialogue evaluation, which aims to evaluate dialogues efficiently and accurately, plays an important role in dialogue systems. On the one hand, it provides a basis for cross-model comparison, on the other hand, it points out the direction for model improvement. While recent efforts are being put into making learning-based evaluation metrics correlate better with human evaluation, robust metrics for parallel corpora and multiple domains remain unexplored. Parallel corpora refer to corpora express the same idea in different ways (e.g., translation, paraphrasing or back-translation). In Dialogue System Technology Challenge 11 (DSTC11)[1], the track4 "Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems" proposes such a challenge, which consists of two sub-tasks. In sub-task1 ("Metrics for multilingual data"), all participants need to develop effective automatic open-ended and multilingual (i.e., English, Spanish and Chinese) dialogue evaluation metrics that perform similarly when evaluated over all the languages. In sub-task2 ("Robust metrics"), all participants need to develop effective automatic open-ended dialogue evaluation metrics that perform robustly when evaluated over back-translated/paraphrased sentences in English. For both tasks, the developed metrics should be correlated to human judgements well and explainable.

Current automatic dialogue evaluation metrics include word overlap-based metrics, embedding-based metrics and learning-based metrics. Word overlap-based metrics (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005)) evaluate candidate response through its overlapping words with reference response. Embedding-based metrics (e.g., Greedy Matching (Rus and Lintean, 2012) and Vector Extrema (Forgues et al., 2014)) firstly obtain sentence representation through word embedding (e.g., word vector (Mikolov et al., 2013)), then, the semantic similarity between candidate response and reference response is calculated by their representation for dialogue evaluation. However, due to the one-to-many nature of open-domain dialogue (Zhao et al., 2017), the two referenced based metrics above have been shown to be poorly correlated with human evaluation (Liu et al., 2017). Learning-based metrics aim to predict the score of certain quality of candidate response and have shown great correlation with human evaluation (Tao et al., 2018). Previous study of learning-based

♠ Equal Contribution.
♡ Corresponding Author.

[1] https://dstc11.dstc.community/home

Proceedings of The Eleventh Dialog System Technology Challenge, pages 123–132
September 11, 2023

metrics have explored topic transition dynamics in dialogue (Huang et al., 2020), composition of fine-grained qualities (Mehri and Eskenazi, 2020b; Phy et al., 2020; Zhang et al., 2022) and quantifiable dialogue coherence evaluation (Ye et al., 2021). However, all of the current metrics are tested on a monolingual setup, and fail to consider the robustness of metrics in noisy settings.

To address the above issues, we propose **P**arallel **C**orpora **A**lignment **F**ramework (PCAF), which improves the consistency and robustness of model evaluation on parallel corpora. Parallel corpora refer to corpora that express the same idea in different manners (e.g., translation, paraphrasing or back-translation). Firstly, parallel corpora are aligned in semantic space through parallel-corpora-aligned contrastive learning. Then, parallel-corpora-aligned knowledge distillation (Hinton et al., 2015) on multiple datasets is applied to improve model's evaluation capability across multiple data domains. Our approach ranks second on the final test data of DSTC11 track4 sub-task1 ("Multilingual Automatic Evaluation Metrics", turn-level) with an average Spearman correlation score of 36.57% and third on the sub-task2 ("Robust Automatic Evaluation Metrics", turn-level) with an average Spearman correlation score of 38.30%.

Our contributions are summarized as follows:

- We propose a novel framework PCAF which improves the consistency and robustness of model evaluation on parallel corpora.

- Our approach ranks second on the final test data of DSTC11 track4 sub-task1 ("Multilingual Automatic Evaluation Metrics", turn-level) and third on the sub-task2 ("Robust Automatic Evaluation Metrics", turn-level), which proves the strong generalization ability and robustness of our proposed approach.

2 Related Work

Automatic dialogue evaluation is of great importance to dialogue systems. It can be divided into dialogue-level and turn-level, while dialogue-level pays attention to the overall evaluation of dialogue system, turn-level mainly evaluate the quality of candidate response according to the provided dialogue history. This paper mainly focus on turn-level automatic dialogue evaluation.

Word overlap-based metrics and embedding-based metrics are standard automatic evaluation metrics (Zhang et al., 2022). However, these metrics which assess dialogue based on reference response have been shown to be inaccurate for dialogue evaluation (Liu et al., 2017). Learning-based metrics was then proposed, which adopts deep learning models and aims to predict human-like scores to input responses (Lowe et al., 2017).

Learning-based metrics can be divided into supervised and self-supervised. Supervised metrics (Lowe et al., 2017) highly depend on human-annotated training data, which are not widely studied due to the lack of such data. Self-supervised metrics utilize human response as positive responses, and negative responses are obtained through negative sampling, thus, positive-negative responses pairs are constructed for model training. As human conversations are readily available (e.g., DailyDialogue (Li et al., 2017)), various self-supervised metrics have been proposed. Grade (Huang et al., 2020) applies graph reasoning to model topic transition dynamics in dialogue. Maude (Sinha et al., 2020) distinguishes the positive and negative responses using NCE loss (Gutmann and Hyvärinen, 2010). USR (Mehri and Eskenazi, 2020b) trains one language model and two dialogue retrieval models to measure five qualities respectively and regresses them to an overall score. MME-CRS (Zhang et al., 2022) trains five sub-models to evaluate five qualities respectively and weighted them to an overall judgement.

However, all of the current metrics are tested on a monolingual setup, and fail to consider metrics' robustness to changes in domain and expression. PCAF can effectively alleviate these problems and shows great generalization ability and robustness.

3 Methodology

Figure 1 illustrates the pipeline of our proposed PCAF, a two-stage training framework which aligns parallel corpora in semantic space and improves model's generalization ability. In this section, we will introduce parallel-corpora-aligned pre-training and parallel-corpora-aligned knowledge distillation for generalization in detail.

3.1 Model Architecture

The metric model consists of a encoder for feature extraction and a predictor for score prediction. Specifically, we adopt XLM-RoBERTa (Conneau et al., 2020) and RoBERTa (Liu et al., 2019) as the encoder network for sub-task1 and sub-task2 re-

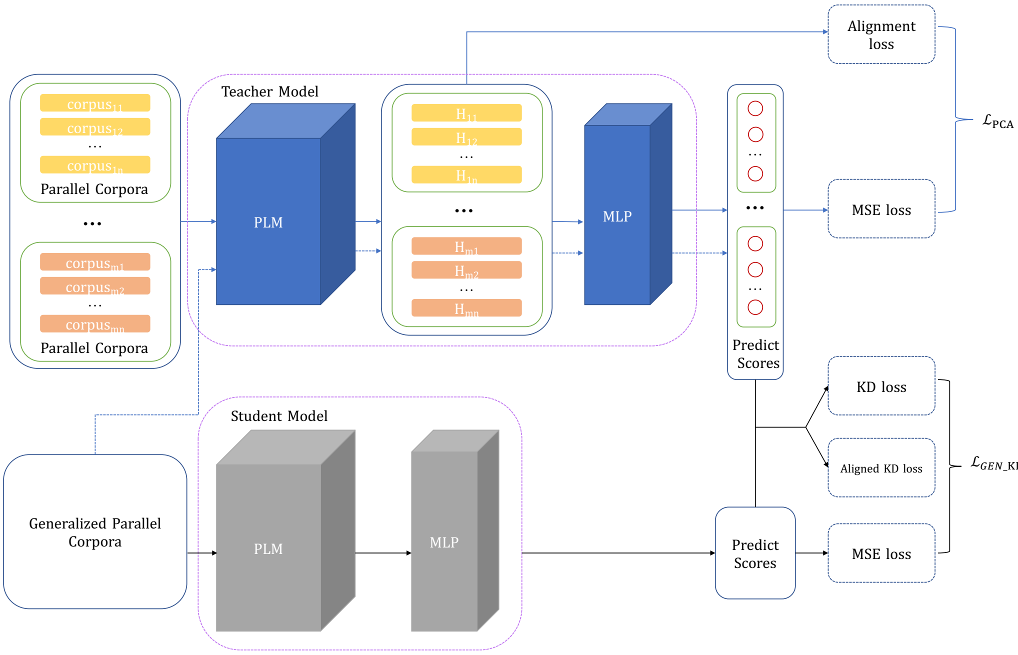


Figure 1: The overall pipeline of our PCAF, including parallel-corpora-aligned pre-training and knowledge distillation on multi-dataset. The solid blue one-way arrows show the process of parallel-corpora-aligned pre-training stage. During pre-training stage, parallel corpora are aligned in semantic space through contrastive learning, and model's prediction capability is optimized by MSE loss. The solid black one-way arrows and dotted blue one-way arrows illustrate the KD stage, where the student model is initialized with the teacher model and optimized by GEN_KD loss to improve model's generalization ability.

spectively, and a three-layer multi-layer perceptron (MLP) is used as the predictor network.

Given the context $\mathbf{c} = \{u_1, u_2, ..., u_{|c|}\}$ and response $\mathbf{r} = \{u_r\}$ where u_i is utterance of context or response, the representation of response is firstly obtained by the pooled output feature of XLM-RoBERTa:

$$v_r = Pooler(Encoder(\mathbf{c}, \mathbf{r})) \tag{1}$$

Then, the quality score of response is predicted by:

$$s = MLP(v_r) \tag{2}$$

3.2 Parallel-Corpora-Aligned Pre-training

Parallel corpora refers to corpora that express the same idea in different ways. Common parallel corpora are utterances written in different languages or their paraphrased, back-translated versions. Aligning parallel corpora in semantic space can not only improve model's evaluation capability in multi-lingual settings, but also make model robust to changes in form of the utterances. Besides, training an aligned model on one language can improve model's performance on another language as well.

Parallel corpora is aligned in semantic space through contrastive learning. Formally, given a training dataset $\mathcal{D}_{pc} = \{\mathcal{P}_i\}, \mathcal{P}_i = \{(c_{ij}, r_{ij}, \overline{r}_{ij})\}$ where c_{ij} and r_{ij} are a ground-truth context-response pair and $\overline{r}_{ij}$ is a negative response sampled by negative sampling strategy. Each dialogue in $\mathcal{P}_i$ except the negative responses expresses the same idea but in different manners , while dialogues in $\mathcal{P}_i$ express different ideas with dialogues in $\mathcal{P}_j$ $(i \neq j)$. Let $v_{ij}, \overline{v}_{ij}$ be the representation of $r_{ij}, \overline{r}_{ij}$, the similarity between all responses is:

$$\begin{aligned} S_{all} = & \sum_{i=1}^{|\mathcal{D}_{pc}|-1} \sum_{j=1}^{|\mathcal{P}_i|} \sum_{m=i+1}^{|\mathcal{D}_{pc}|} \sum_{n=1}^{|\mathcal{P}_m|} exp(cos\_sim(v_{ij}, v_{mn})/\tau) \\ & + \sum_{i=1}^{|\mathcal{D}_{pc}|-1} \sum_{j=1}^{|\mathcal{P}_i|} \sum_{m=i+1}^{|\mathcal{D}_{pc}|} \sum_{n=1}^{|\mathcal{P}_m|} exp(cos\_sim(\overline{v}_{ij}, v_{mn})/\tau) \\ & + \sum_{i=1}^{|\mathcal{D}_{pc}|-1} \sum_{j=1}^{|\mathcal{P}_i|} \sum_{m=i+1}^{|\mathcal{D}_{pc}|} \sum_{n=1}^{|\mathcal{P}_m|} exp(cos\_sim(\overline{v}_{ij}, \overline{v}_{mn})/\tau) \end{aligned} \tag{3}$$

the similarity between responses in $\mathcal{P}_i$ is:

Dataset	Spanish Translation	Chinese Translation	Paraphrases	English Back-translation
DBDC (Higashinaka et al., 2016)	✓	-	✓	✓
CMU_DoG (Zhou et al., 2018)	✓	-	✓	✓
Cornell Movie-Dialogs (Danescu-Niculescu-Mizil and Lee, 2011)	-	✓	✓	✓
DailyDialog (Li et al., 2017)	✓	✓	✓	✓
DECODE (Nie et al., 2021)	✓	-	✓	✓
EmotionLines (Hsu et al., 2018)	✓	-	✓	✓
EmpathicDialogues (Rashkin et al., 2019)	✓	✓	✓	✓
Holl-E (Moghe et al., 2018)	✓	-	✓	✓
MEENA (Adiwardana et al., 2020)	✓	-	✓	✓
MELD (Poria et al., 2019)	✓	-	✓	✓
MetalWOz (Lee et al., 2019)	✓	-	✓	✓
Movie-DiC (Banchs, 2012)	✓	-	✓	✓
PersonaChat (Zhang et al., 2018)	✓	✓	✓	✓
SentimentLIAR (Upadhayay and Behzadan, 2020)	✓	-	✓	✓
Switchboard Coherence (Cervone and Riccardi, 2020)	-	✓	✓	✓
Topical-Chat (Gopalakrishnan et al., 2019)	✓	✓	✓	✓
Wizard of Wikipedia (Dinan et al., 2019)	✓	✓	✓	✓
WOCHAT (D'Haro et al., 2016)	✓	-	✓	✓

Table 1: Training sets provided by DSTC11 Track4 organizers. The source language of these datasets is English, and all of them are provided with English back-translation and paraphrases.

$$S_{\mathcal{P}_i} = \sum_{j=1}^{|\mathcal{P}_i|-1} \sum_{k=j+1}^{|\mathcal{P}_i|} exp(cos\_sim(v_{ij}, v_{ik})/\tau) \quad (4)$$

and the alignment loss is:

$$l_{align} = -\sum_{i=1}^{|\mathcal{D}_{pc}|} log \frac{S_{\mathcal{P}_i}}{S_{all}} \quad (5)$$

together with the MSE loss:

$$l_{mse} = \sum_{i=1}^{\mathcal{D}_{pc}} \sum_{j=1}^{\mathcal{P}_i} ((1 - s_{ij})^2 + \overline{s}_{ij}^2) \quad (6)$$

the final loss of parallel-corpora-align pre-training is:

$$\mathcal{L}_{PCA} = l_{align} + l_{mse} \quad (7)$$

3.3 Knowledge Distillation on Multiple Datasets

After parallel-corpora-aligned pre-training, the model M is further trained by parallel-corpora-aligned knowledge distillation on multiple datasets to attain a better generalization ability.

Given parallel corpora $\mathcal{P} = (c_i, r_i, \overline{r}_i)$ where r_i and $\overline{r}_i$ are positive-negative response pair, M_t is the teacher model, and M_s is the student model which is initialized with the teacher model. Let the teacher model and student model predict the score of the response pair respectively, and get $s_{ti}, \overline{s}_{ti}, s_{si}, \overline{s}_{si}$.

The student model is firstly optimized by MSE loss:

$$l_i^{kd\_mse} = (1 - s_{si})^2 + \overline{s}_{si}^2 \quad (8)$$

Then, we utilize the teacher model's predictions as soft targets. Besides, considering the parallel corpora all express the same idea, we take the average of teacher model's predictions of positive responses as the label for the entire parallel corpora's positive responses, and the KD loss is formulated as:

$$l_i^{kd} = (s_{ti} - s_{si})^2 + (\overline{s}_{ti} - \overline{s}_{si})^2 + (\frac{\sum_{k=1}^{|\mathcal{P}|} s_{ti}}{|\mathcal{P}|} - s_{si})^2 \quad (9)$$

The overall loss in KD stage is the weighted sum of $l_i^{kd\_mse}$ and l_i^{kd}:

$$\mathcal{L}_{GEN\_KD} = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} (\alpha * l_i^{kd\_mse} + (1 - alpha) * l_i^{kd}) \quad (10)$$

where α is the hyperparameter.

4 Experiments

4.1 Datasets

As shown in Table 1, the organizers of DSTC11 Track4 provide 18 human-human dialogue datasets as training set. Table 2 shows the result of our preliminary experiment of datasets comparison. For each dataset, we randomly sampled 3k data to train a BERT + MLP model which is further tested on the provided development sets respectively. As the model trained on DailyDialog(Li et al., 2017) shows the highest Spearman correlation among the 18 human-human dialogue datasets, we select DailyDialog as the pre-training dataset. Besides, SentimentLIAR and Switchboard Coherence are excluded as they encountered training collapse in our preliminary experiment.

Dataset	Spearman(%)	rank
DBDC	21.41	15
CMU_DoG	24.31	13
Cornell Movie-Dialogs	27.26	9
DailyDialog	31.48	1
DECODE	27.43	7
EmotionLines	30.25	2
EmpatheticDialog	28.37	6
Holl-E	24.00	14
MEENA	24.44	12
MELD	27.37	8
MetalWOz	25.28	11
Movie-DiC	28.81	5
PersonaChat	25.55	10
SentimentLIAR	-	-
Switchboard Coherence	-	-
Topical-Chat	29.98	3
Wizard of Wikipedia	29.11	4
WOCHAT	29.11	16

Table 2: Result of preliminary experiment of datasets comparison. We test the BERT + MLP model trained with 3k English data from each dataset respectively on provided development sets. The Spearman correlation is the average result of all development sets.

As for the development set, the organizers provide the following 14 turn-level datasets which have been automatically translated to Spanish and Chinese, and back-translated to English:

- CONVAI2-GRADE (CG) (Huang et al., 2020)
- DAILYDIALOG-GRADE (DH) (Huang et al., 2020)
- DAILYDIALOG-GUPTA (DG) (Gupta et al., 2019)
- DAILYDIALOG-ZHAO (DZ) (Zhao et al., 2020)
- DSTC7 (D7) (Galley et al., 2019)
- EMPATHETIC-GRADE (EG) (Huang et al., 2020)
- FED-TURN (FT) (Mehri and Eskenazi, 2020a)
- HUMOD (HM) (Merdivan et al., 2020)
- PERSONA-USR (PU) (Mehri and Eskenazi, 2020b)
- PERSONA-ZHAO (PZ) (Zhao et al., 2020)
- TOPICAL-USR (TU) (Mehri and Eskenazi, 2020b)

Team	EN	ZH	ES	Multilingual AVG
Deep AM-FM	29.40	7.53	18.26	18.40
TOP 1	48.18	39.36	58.90	48.81
TOP 2 (**ours**)	22.14	31.12	56.44	36.57
TOP 3	37.02	7.01	19.83	21.29
TOP 4	14.69	10.54	8.08	11.10

Table 3: The Spearman correlation (%) of baseline Deep AM-FM and top 4 teams on the test datasets of sub-task1 (turn-level). Only the best result of each team is shown in the table.

Team	Robust AVG
Deep AM-FM	0.3387
TOP 1	0.4890
TOP 2	0.4190
TOP 3 (**ours**)	0.3833
TOP 4	0.2697

Table 4: The Spearman correlation (%) of baseline Deep AM-FM and top 4 teams on the test datasets of sub-task2 (turn-level). Only the best result of each team is shown in the table.

- JSALT (JS) (Zhang et al., 2021)
- CHATEVAL (CS) (Sedoc et al., 2019)
- DSTC10 (D10) (Zhang et al., 2021)

Considering the multilingual setting of sub-task1, model M_1 is only trained on DailyDialog, EmpatheticDialog, PersonaChat, Topical-Chat and Wizard of Wikipia, which are translated into both Chinese and Spanish. While DailyDialog is used as the pre-training dataset, all of the 5 datasets above take part in the knowledge distillation stage of PCAF.

For sub-task2, all of the training sets except for SentimentLIAR and Switchboard Coherence are used to train model M_2. Still, DailyDialog is used as the pre-training dataset and all of these datasets above take part in the knowledge distillation stage of PCAF.

4.2 Implementation Details

In sub-task1, we adopt XLM-RoBERTa-Large as the encoder, and the parallel corpora is English-Chinese-Spanish corpora.

In sub-task2, we adopt RoBERTa-Large as the encoder, and the parallel corpora is English-Paraphrases corpora.

For both of the two tasks, we adopt Adam (Kingma and Ba, 2017) as the optimizer and set

Model	Language	Appropriateness	Content Richness	Grammatical Correctness	Relevance	Average
Deep AM-FM	EN	34.32	31.03	19.37	32.90	29.40
Deep AM-FM	ZH	12.32	14.18	3.61	0.02	7.53
Deep AM-FM	EZ	0.94	2.36	32.97	36.79	18.26
PCAF	EN	15.53	57.28	2.52	17.22	23.14
PCAF	ZH	23.18	49.56	7.43	46.35	31.63
PCAF	ES	53.48	77.86	36.87	57.57	56.44

Table 5: Fine-grained result of Deep AM-FM and our best submission on the test datasets of sub-task1 (turn-level).

Model	Coherence	Engageness	Informativeness	Overall	Average
Deep AM-FM	29.37	37.91	30.66	37.54	33.87
PCAF	39.66	42.45	28.34	42.87	38.33

Table 6: Fine-grained result of Deep AM-FM and our best submission on the test datasets of sub-task2 (turn-level).

batchsize as 32, learning rate as 5e-6, τ as 0.05, α as 0.2, and the model is trained on one single RTX 3090. Besides, epochs of the pre-training stage and KD stage are both set as 10.

4.3 Comparison Result

According to DSTC11 Track4, the turn-level metrics are evaluated by the following dimensions in both sub-task1 and sub-task2:

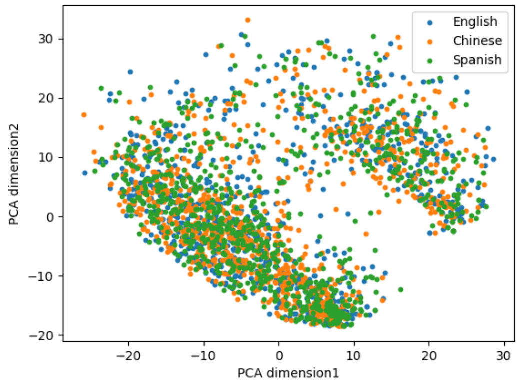
- **Appropriateness** - The response is appropriate given the preceding dialogue.
- **Content Richness** - The response is informative, with long sentences including multiple entities and conceptual or emotional words.
- **Grammatical Correctness** - Responses are free of grammatical and semantic errors.
- **Relevance** - Responses are on-topic with the immediate dialogue history.

For each submission, Spearman correlation at dimension-level will be calculated separately for each task. Then, the Spearman correlation scores obtained will be averaged. Finally, the Spearman correlation scores will be ranked.

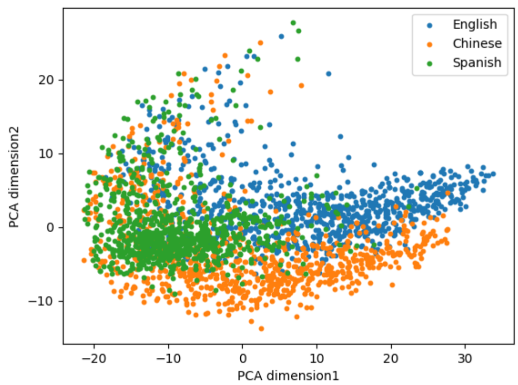
We compare our approach with Deep AM-FM (Zhang et al., 2020) and the top 4 teams in sub-task1 and sub-task2 in Table 3 and Table 4 respectively, and the fine-grained results are reported in Table 5 and Table 6. PCAF ranks second and third in sub-task1 and sub-task2 in the comparison with Deep AM-FM and other teams' approaches, showing the effectiveness of our approach.

4.4 Ablation Studies

To verify the contribution of parallel-corpora-aligned pre-training and parallel-corpora-aligned



(a) aligned



(b) unaligned

Figure 2: PCA results of metric model M trained with (top) and without (bottom) alignment loss on DailyDialog dataset

Metric	EN	ZH	ES	Multilingual AVG
PCAF	42.26 ± 0.001	40.49 ± 0.001	39.75 ± 0.002	40.83 ± 0.001
kd on DailyDialog	40.94 ± 0.001	38.76 ± 0.001	38.80 ± 0.001	39.50 ± 0.001
w/o kd	39.42 ± 0.009	37.86 ± 0.007	37.13 ± 0.007	38.13 ± 0.007
w/o kd & align	37.70 ± 0.030	37.35 ± 0.007	33.39 ± 0.054	36.15 ± 0.029

Table 7: The Spearman correlation (%) of PCAF ablation study. KD on DailyDialog means the knowledge distillation is only apply to DailyDialog training set. W/o align means the alignment loss is not involved in pre-training stage. Standard deviations are presented in gray color.

Corpora	Original	Paraphrase	English Back-translation
ori	36.67 ± 0.018	27.80 ± 0.018	32.96 ± 0.022
para	37.30 ± 0.011	31.63 ± 0.005	36.91 ± 0.006
bt	32.84 ± 0.016	27.07 ± 0.011	30.64 ± 0.014
ori + para	39.44 ± 0.001	32.34 ± 0.003	37.25 ± 0.011
ori + bt	34.62 ± 0.022	25.53 ± 0.018	26.57 ± 0.022
ori + para + bt	38.85 ± 0.007	32.55 ± 0.007	35.86 ± 0.021
ori + para + kd	42.33 ± 0.001	34.05 ± 0.001	39.47 ± 0.002

Table 8: Ablation study of different corpora combination, where ori, para, bt, kd stands for original utterances, paraphrases, back-translation and knowledge distillation respectively. Models are tested on the original, paraphrases and English back-translation corpora of development sets by Spearman correlation (%) respectively. Standard deviations are presented in gray color.

knowledge distillation on multiple datasets, we further conduct ablation studies on the provided development sets.

Table 7 shows the results of ablation study on sub-methods of PCAF. According to the results, both parallel-corpora-aligned pre-training and parallel-corpora-aligned knowledge distillation make contribution to the improvement of the model's performance. The alignment loss not only improves the evaluation ability of the model, but also improves the stability of the model training according to the standard deviation of model's validation results. We further visualize the encoded features on DailyDialog through Principal Component Analysis (PCA). As shown in Figure 2, compared to models trained without alignment loss, model trained with alignment loss has a more compact feature distribution on parallel corpora for the same sequence, showing that alignment loss effectively aligns model's representation of parallel corpora. We suppose that, as the representation of parallel corpora is pre-aligned in multilingual language model, the absence of alignment loss during pre-training may disturb model's original multilingual-aligned knowledge, which is shown in Figure 2(b). Besides, knowledge distillation is an important stage of PCAF, and the comparison between kd single DailyDialog and kd on five datasets shows that kd on multiple datasets do improves the generalization ability of model.

Training data of different parallel corpora combination of sub-task2 is also explored, Table 8 shows the result of this experiment. The combination of ori + para achieves the highest performance of the provided development sets, while the English back-translation corpora always degrades the performance of the model. The reason of such phenomena is unclear at present, and we leave it to our future work.

5 Conclusion

In this paper, we propose PCAF, a parallel-corpora-aligned training framework for training multilingual and robust turn-level automatic dialogue evaluation metrics. PCAF treats corpora express the same idea in different ways as parallel-corpora, which is aligned during both PCAF pre-training stage and PCAF knowledge distillation stage. Experiment results show that PCAF achieves a great performance, which demonstrates the effectiveness of PCAF. The effectiveness of each sub-method of PCAF is also proved through ablation study.

6 Limitations

DSTC11 Task4 requires the proposed metrics to evaluate the dialogues on multiple fine-grained qualities. However, we only train the metric model to evaluate the appropriateness of dialogues, whose results are further used as the evaluation results of other qualities. As PCAF can be integrated into the training process under any parallel-corpora setting, we can further try to train the model to evaluate other fine-grained qualities of dialogues with PCAF.

Besides, despite the DSTC11 Task4 organizers allow the participants to fine-tune their system over a subset of the development data, our submitted model is not fine-tuned with those human-annotated datasets. While we train our metric model under self-supervised learning framework, fine-tuning it on supervised datasets may improve models evaluation performance, which will be explored in our future work.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Rafael E. Banchs. 2012. Movie-DiC: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–207, Jeju Island, Korea. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this dialogue coherent? learning from dialogue acts and entities.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Luis F D'Haro, Bayan Abu Shawar, and Zhou Yu. 2016. Rewochat 2016–shared task description report. In *Proceedings of the workshop on collecting and generating resources for chatbots and conversational agents-development and evaluation (RE-WOCHAT)*, page 39.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3146–3150, Portorož, Slovenia. European Language Resources Association (ELRA).

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

S Lee, H Schulz, A Atkinson, J Gao, K Suleman, L El Asri, M Adada, M Huang, S Sharma, W Tay, et al. 2019. Multi-domain task-completion dialog challenge. *Dialog system technology challenges*, 8(9).

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2017. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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 1699–1713, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, NAACL HLT '12, page 157–162, USA. Association for Computational Linguistics.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Bibek Upadhayay and Vahid Behzadan. 2020. Sentimental liar: Extended corpus and deep learning models for fake claim classification. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. 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 2718–2729, Online. Association for Computational Linguistics.

Chen Zhang, Luis D'Haro, Rafael Banchs, Thomas Friedrichs, and Haizhou Li. 2020. *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation*, pages 53–69.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems.

Pengfei Zhang, Xiaohui Hu, Kaidong Yu, Jian Wang, Song Han, Cao Liu, and Chunyang Yuan. 2022. Mme-crs: Multi-metric evaluation based on correlation re-scaling for evaluating open-domain dialogue.

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, Melbourne, Australia. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).