Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 146–153
Marseille, 20-25 June 2022

An Empirical Study on the Overlapping Problem of Open-Domain Dialogue Datasets

Yuqiao Wen, Guoqing Luo, Lili Mou
Dept. Computing Science, Alberta Machine Intelligent Institute (Amii), University of Alberta
yuqiao@ualberta.ca, gluo@ualberta.ca, doublepower.mou@gmail.com

Abstract

Open-domain dialogue systems aim to converse with humans through text, and dialogue research has heavily relied on benchmark datasets. In this work, we observe the overlapping problem in DailyDialog and OpenSubtitles, two popular open-domain dialogue benchmark datasets. Our systematic analysis then shows that such overlapping can be exploited to obtain fake state-of-the-art performance. Finally, we address this issue by cleaning these datasets and setting up a proper data processing procedure for future research.

Keywords: Open-domain dialogue, Data cleaning

1. Introduction

Open-domain dialogue generation is the task of generating natural language utterances to converse with humans (Shang et al., 2015; Mou et al., 2016; Tian et al., 2017). Such systems have wide applications in the industry. For example, XiaoIce[1] from Microsoft has been deployed to more than 40 platforms and has gained 660 million active users since its launch in 2014 (Zhou et al., 2020).

There have been several benchmark datasets for open-domain dialogue generation, and they are largely advancing the field. For example, DailyDialog (Li et al., 2017) and OpenSubtitles (Lison et al., 2018) have been extensively used in recent studies (Cai et al., 2020a; Sun et al., 2021; Zhou et al., 2021; Wang et al., 2021).

However, we observe a common problem of existing open-domain dialogue datasets: the training and test sets tend to overlap with each other. That is, a large number of identical or near-identical dialogues appear in both training and test sets, probably due to mistakes in data collection and preprocessing.

We further observe that such overlaps cause bizarre behaviors in training dialogue systems:

1. Common evaluation metrics are heavily inflated. A dialogue system can achieve perfect test performance on overlapping samples by memorizing the training set. However, such high performance is fake, as it reflects the dialogue system's ability to memorize rather than its conversational skills.
2. Reported performance based on overlapping samples is arbitrary. Due to overlapping, performance may continue to grow even after 1000 epochs as the dialogue system continues to memorize more training samples. Since most researchers do not train their models long enough (e.g., for 1000 epochs), their reported performance is arbitrary depending on the maximum epoch or the early stopping strategy, making the comparison of state-of-the-art models meaningless.
3. Generated utterances are over-informative. For example, we observe that models trained on overlapping samples can predict the speaker's name correctly with no context. Such behaviour is not realistic for any dialogue systems, highlighting the issues with overlapping datasets.

Therefore, we argue that it is crucial to revisit and clean benchmark dialogue datasets for rigorous scientific research. Our contributions are threefold. First, we observe and report the overlap issue of existing dialogue benchmark datasets. Second, we perform systematic analyses to show the consequences of such overlapping. Third, we provide cleaned datasets[2] for future open-domain dialogue research.

In light of our research, we advocate the following practice in future open-domain dialogue research:

1. Avoid comparing state-of-the-art models on overlapping datasets; and
2. Always revisit the quality of existing and future datasets for dialogue research.

2. Related Work

Dialogue systems can be categorized into two main paradigms: task-oriented and open-domain. Task-oriented dialogue systems aim to achieve specific tasks such as finding restaurants and booking hotels. For example, the ATIS corpus (Hemphill et al., 1990) is an early task-oriented dataset that focuses on air travel, collected through the Wizard-of-Oz (Kelley, 1984) scheme. Recently, Wen et al. (2017) adopt Wizard-of-Oz to crowd-sourcing and largely reduce the cost for collecting large annotated data. Following their approach, many more datasets are collected,

[1] `http://www.xiaoice.com`

[2] Our cleaned datasets and source code are available at: `https://github.com/yq-wen/overlapping-datasets`

such as the Frames corpus (El Asri et al., 2017), MultiWOZ (Budzianowski et al., 2018), and CrossWOZ (Zhu et al., 2020).

On the other hand, open-domain dialogue systems aim to hold engaging and open-ended conversations with humans. Early systems such as ELIZA (Weizenbaum, 1966) and ALICE (Wallace, 2009) are based on manually crafted rules. Recent advances in deep neural networks make it possible to train dialogue systems end-to-end using massive dialogue corpora (Shang et al., 2015; Li et al., 2016; Serban et al., 2016). Therefore, researchers have created many open-domain dialogue datasets, such as DailyDialog (Li et al., 2017), Persona-Chat (Zhang et al., 2018), and Topical-Chat (Gopalakrishnan et al., 2019).

While benchmarking datasets largely advance the NLP field over the past decades, it is not uncommon that benchmarked datasets have flaws. For example, the task-oriented dialogue dataset ATIS is adapted to the semantic parsing task. As pointed out by Guo et al. (2020) and Huang et al. (2021), a large number of samples become identical when researchers anonymize the entities in an utterance (Dong and Lapata, 2016). Schumann et al. (2020) identify a problem in the summarization task that the previous benchmark setting does not properly enforce summary length, allowing "state-of-the-art" models to gain performance by generating over-lengthed summaries. These highlight the importance of properly benchmarking a task for NLP research.

The overlapping problem of DailyDialog was first reported by the previous work of one of our coauthors (Bahuleyan et al., 2019; Khan et al., 2020). However, the community has not been adequately aware of this. Our new paper studies the problem more systematically. We show that the problem is more severe than we thought as it occurs in different datasets. We perform more detailed analyses and provide an approach to deduplicate overlapping samples.

3. Bizarre Behaviors When Data Overlap

We present our empirical findings for the overlapping problem and its consequences with two commonly used open-domain dialogue datasets: DailyDialog[3] and OpenSubtitles[4]. While overlapping samples exist in many datasets, we find the problem most severe in DailyDialog and OpenSubtitles.

Overlapping Statistics. We represent an utterance as a bag of words, and compute the overlap ratio between two utterances $\mathbf{u} = \{u_1, \cdots, u_m\}$ and $\mathbf{v} = \{v_1, \cdots, v_n\}$ as $R(\mathbf{u}, \mathbf{v}) = \frac{2|\mathbf{u} \cap \mathbf{v}|}{|\mathbf{u}|+|\mathbf{v}|}$. A data sample is a tuple $\mathbf{x} = (\mathbf{c}, \mathbf{r})$, where $\mathbf{c}$ is the context and $\mathbf{r}$ is the response. We then compute the overlap ratio of two samples $\mathbf{x} = (\mathbf{c}, \mathbf{r})$ and $\mathbf{x}' = (\mathbf{c}', \mathbf{r}')$ as $R(\mathbf{x}, \mathbf{x}') = \min\{R(\mathbf{c}, \mathbf{c}'), R(\mathbf{r}, \mathbf{r}')\}$; the min operator rules out false overlapping caused by generic utterances (e.g., *hello*), whose responses may be different. Finally, the overlap ratio of a test sample $\mathbf{x}$ against the training dataset $\mathcal{D}_{\text{train}}$ is given by $R(\mathbf{x}, \mathcal{D}_{\text{train}}) = \max_{\mathbf{x}' \in \mathcal{D}_{\text{train}}} R(\mathbf{x}, \mathbf{x}')$.

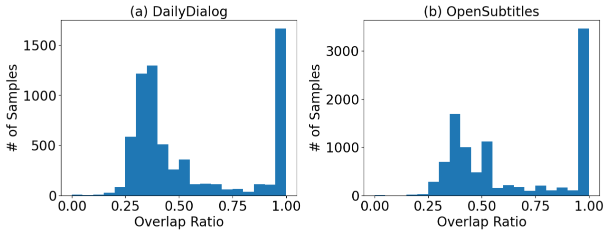

Figure 1: Overlap histogram of test samples against the training set on (a) DailyDialog and (b) OpenSubtitles

Figure 1 shows the histogram of the overlap ratio in DailyDialog and OpenSubtitles. We find that 23.15% and 34.49% test samples are identical to training samples in the two datasets, respectively. Further, many samples are near-identical. In Table 1, for instance, a test sample only contains additional speaker information ("A ::" and "B ::") compared with another training sample, resulting in an overlap ratio of 0.80.

As shown, such overlapping does not naturally arise from human conversations; rather, they are caused by oversights in data collection and preprocessing. For DailyDialog, the dataset is constructed by crawling English learning websites (Li et al., 2017), and overlapping samples probably come from the common learning materials shared among different websites. For OpenSubtitles, Lison et al. (2018) group subtitles based on their IMDb identifiers. However, we find that the same movie may correspond to different identifiers if it has different versions. For example, the South Korean movie My Sassy Girl[5] and its American remake[6] contain highly overlapping dialogues, but they are treated as different movies based on their IMDb identifiers. Such overlapping raises serious concerns on whether these datasets are appropriate for benchmarking open-domain dialogue research.

Learning Curves. We observe that overlapping samples have an undesired effect on training dialogue systems. We compare the learning curves on the original test set and a deduplicated test set, where we remove samples with an overlap ratio of greater than 0.80. In terms of the neural architecture, we fine-tune a T5-small model (Raffel et al., 2020) here.

Figure 2 shows the learning curves in terms of the

[3] `http://yanran.li/files/ijcnlp_dailydialog.zip`, accessed on Oct 27, 2021.

[4] Since the original OpenSubtitles dataset does not provide data splits and there is no standard practice, we use the splits from a recent study (Wang et al., 2021) as a representative: `https://drive.google.com/file/d/1U4M0h9tLNeCyu9JBfSgR3r5EE6IIqyNZ/`, accessed on Nov 8, 2021

[5] `https://www.imdb.com/title/tt0293715/`

[6] `https://www.imdb.com/title/tt0404254/`

0.60	Train	Context	Do you have a fever ?
		Response	I don't know , but I feel terrible .
	Test	Context	Do you have an airsickness ?
		Response	I don't know . But I have a carsickness .
0.80	Train	Context	Nice to meet you , Mr . Wilson .
		Response	Tim , please . Please be seated .
	Test	Context	B :: Nice to meet you , Mr . Wilson .
		Response	A :: Tim , please . Please be seated .
1.00	Train	Context	It seldom rains this summer .
		Response	Yeah , some places are very short of water .
	Test	Context	It seldom rains this summer .
		Response	Yeah , some places are very short of water .

Table 1: Training and test samples with their corresponding overlap ratios from the original DailyDialog dataset.

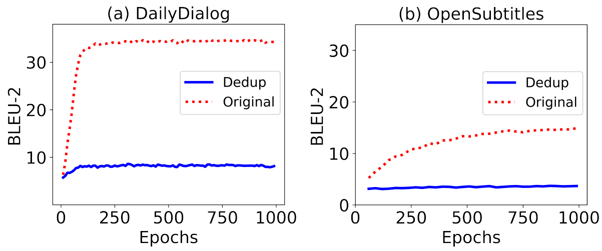

Figure 2: BLEU-2 learning curves for the original dataset and the deduplicated dataset for (a) DailyDialog and (b) OpenSubtitles. Samples with an overlap greater than 0.80 are considered duplicates and are removed for the deduplicated dataset.

BLEU-2 metric. On DailyDialog, the model achieves ∼34 BLEU-2 with the original test set; however, the same model only achieves ∼8 BLEU-2 after deduplication. The trend is similar on OpenSubtitles: 15 BLEU-2 with the original test set versus 3 BLEU-2 with the deduplicated one. The results suggest that overlapping samples heavily inflate BLEU scores, and that the high performances of the alleged "state-of-the-art" models mainly come from memorizing training samples.

In addition, we observe that it takes more epochs for the test set performance to converge when the test samples overlap with the training set. For example, the BLEU-2 score still increases after 1000 epochs on OpenSubtitles, which contains more complex and diverse conversations than the DailyDialog dataset. Since most researchers do not train their models for 1000 epochs, their reported performance may be arbitrary, depending on where the model ends up along the learning curve: Wang et al. (2021) report 9.8 BLEU-2 on OpenSubtitles, whereas Sun et al. (2021) report 32.6. This highlights the inconsistency of reported performances.

Overly Informative Outputs. Table 2 illustrates another bizarre behavior that the model generates overly informative outputs. For example, we consider fine-tuning T5-small for single-turn conversations. Given the input *Nice to see you, Patrick*, the model generates *Bob! I hear your team won the match*. The output precisely matches the test reference, achieving a BLEU score of 100. Such an output is overly informative because it is extremely unlikely that the model can correctly predict the speaker's name without any context.

In summary, our qualitative and quantitative analyses show that it is fundamentally flawed to evaluate open-domain dialogue systems on overlapping datasets, which are unfortunately commonly used (and benchmarked) in current research.

4. Dataset Cleaning

We make efforts to clean existing datasets and present results for standard and state-of-the-art models on the cleaned corpora.

In the literature, both single-turn and multi-turn settings are common (Cai et al., 2020a; Zhou et al., 2021; Wang et al., 2021; Gu et al., 2021). We propose to deduplicate data samples in the multi-turn setting, i.e., a dialogue session in DailyDialog and an entire movie in OpenSubtitles are treated as a unit for deduplication. Such a unit can be further split into single turns, so our treatment unifies both settings.

Further, we propose to re-split the training, validation, and test sets of the original corpora during deduplication. If we simply remove the duplicate samples as in Section 3, then at least one of the validation and test sets will be heavily shrunk due to massive overlapping samples, making evaluation noisy and unreliable.

Let us consider two units $\mathbf{u}$ and $\mathbf{v}$ (dialogue sessions or movies) for deduplication. Their overlap ratio is defined as

$$R(\mathbf{u}, \mathbf{v}) = \frac{2|\mathbf{u} \cap \mathbf{v}|}{|\mathbf{u}| + |\mathbf{v}|} \qquad (1)$$

The overlap ratio of a unit $\mathbf{u}$ is then computed as the maximum overlap against all other units in the dataset $\mathcal{D}$:

$$R(\mathbf{u}, \mathcal{D}) = \max_{\mathbf{u}' \in \mathcal{D} \setminus \{\mathbf{u}\}} R(\mathbf{u}', \mathbf{u}) \qquad (2)$$

Note that the above overlap ratio is similar to the one in Section 3 (Overlapping Statistics). However, Eqn. (2) considers the ratio of a deduplication unit (an entire dialogue session or movie), whereas Section 3 considers the ratio of a single-turn conversation. Thus, we do not have the min operator here. Further, the ratio of a unit is computed against the rest of the corpus, whereas

DailyDialog	
Train Input	Nice to see you , Patrick .
Train Ref	Bob ! I hear your team won the match .
Test Input	Nice to see you , Patrick .
Test Ref	Bob ! I hear your team won the match .
Model Output	Bob ! I hear your team won the match .
OpenSubtitles	
Train Input	But you have some strength in you , my dear Hobbit .
Train Ref	What happened, Gandalf ?
Test Input	But you have some strength in you , my dear Hobbit .
Test Ref	What happened, Gandalf ?
Model Output	What happened, Gandalf ?

Table 2: Over-informative model outputs with DailyDialog and OpenSubtitles samples.

Dataset	Version	Context History	# Train	# Validation	# Test
DailyDialog	Li et al. (2017)	Single-Turn	76,052	7,069	6,740
		Multi-Turn	76,052	7,069	6,740
	Cleaned	Single-Turn	60,005	6,594	6,955
		Multi-Turn	60,138	6,612	6,980
OpenSubtitles	Wang et al. (2021)	Single-Turn	1,144,949	20,000	10,000
	Cleaned	Single-Turn	979,230	11,982	12,152
		Multi-Turn	1,002,026	12,289	12,506

Table 3: Data statistics for DailyDialog and OpenSubtitles.

Section 3 computes the ratio of a test sample against the training set. Thus, we have the set minus operator in Eqn. (2).

Then, we iterate through all deduplication units in the corpus. If a unit's overlap ratio exceeds a threshold, then we remove the unit but keep the one that it overlaps the most with. Since there may be more than two units overlapping with each other, we need to repeat the procedure multiple times. Specifically, we recompute the overlap ratios after each iteration through the dataset, and remove additional duplicate samples until convergence.

After that, we split the deduplicated units into training, validation, and test sets. While our deduplication unit is an entire dialogue session or a movie, the resulting corpus can serve for the settings of single-turn and fixed-turn context. We simply split a unit into multiple tuples of contexts and responses. This may result in (a small number of) additional duplicate samples due to the generic conversations, such as *–Thank you. –You're welcome*. Therefore, we further remove exactly overlapping context–response pairs.

Table 3 shows the data statistics for our cleaned datasets for both single-turn and multi-turn settings. Specifically, we follow previous work and use three turns for the multi-turn setting (Li et al., 2017; Cai et al., 2020b; Wang et al., 2021). Our data cleaning strategy enables us to control the size of validation and test sets. For DailyDialog, we keep the size to be similar to the original dataset, which cannot be achieved by a naïve removal of validation/test samples. For OpenSubtitles, the corpus contains much more samples than needed for training a dialogue system; therefore, we follow previous work (Du et al., 2018; Cai et al., 2020a; Wang et al., 2021) and sample a similar number of training, validation, and test samples.

5. Model Performance

In this section, we test multiple standard and state-of-the-art models on our cleaned datasets. We will first show the metrics and models. Then, we will present the experimental results.

5.1. Metrics

We use the BLEU score (Papineni et al., 2002) as an automatic metric to evaluate the quality of the generated responses, following most previous work (Cai et al., 2020a; Sun et al., 2021; Wang et al., 2021). The BLEU metric measures the lexical overlap between the model output and the reference. Note that our BLEU-n measures the geometric mean of i-gram overlap for $i = 1, \cdots, n$.

Additionally, we adopt the Dist metric (Li et al., 2016) to measure the diversity of the generated responses, which is another commonly used metric in dialogue research (Xu et al., 2017; Zhou et al., 2021; Wang et al., 2021). Specifically, Dist-n measures the percentage of distinct n-grams among all generated responses.

5.2. Models

We evaluate multiple standard and state-of-the-art models on our proposed datasets.

- **LSTM with Attention.** We include the long short-term memory network (Hochreiter and Schmidhuber, 1997) with an attention mechanism (Bahdanau et al., 2015) as a baseline. This is a standard model before the Transformer era.
- **Transformer.** In current research, the Transformer (Vaswani et al., 2017) is the most commonly used architecture. It replaces LSTM's recurrent connections with a multi-head attention mechanism and achieves superior performance in different NLP tasks. In this baseline, the Transformer is not pretrained.
- **GPT-2.** The GPT-2 model (Radford et al., 2019) adopts the Transformer architecture, but is pretrained on massive unlabeled corpora. Pretraining is shown to benefit various downstream tasks.
- **T5-small.** The T5 model (Raffel et al., 2020) also uses the Transformer architecture, but works in an encoder–decoder fashion. It is pretrained on a number of text generation tasks, such as translation and summarization. We adopt the T5-small version in our experiments.
- **AdaLabel.** Being one of the recent state-of-the-art models, AdaLabel (Wang et al., 2021) also uses the Transformer encoder–decoder architecture. Instead of the standard cross-entropy training, the model is trained with the adaptive label smoothing technique, where the one-hot labels are smoothed by soft predictions generated from an auxiliary decoder.
- **DialogBERT.** DialogBERT (Gu et al., 2021) is another recent state-of-the-art model. Instead of using a standard Transformer encoder, it uses a hierarchical BERT (Devlin et al., 2019) encoder trained with masked utterance regression and dis-

Context History	Model	Cleaned DailyDialog				Cleaned OpenSubtitles			
		BLEU-2	BLEU-4	Dist-1	Dist-2	BLEU-2	BLEU-4	Dist-1	Dist-2
Single-Turn	LSTM w/ attn	6.56	2.11	3.40	23.50	5.31	1.41	3.10	14.94
	Transformer	7.33	2.56	4.16	25.44	4.89	1.29	3.05	13.88
	T5-small	8.74	3.39	4.63	25.43	6.76	2.07	2.78	8.87
	GPT-2	7.67	2.78	5.38	29.15	7.02	2.15	2.98	11.37
	AdaLabel	6.72	2.29	4.35	26.97	5.66	1.45	3.86	15.33
	DialogBERT†	5.42	2.16	2.57	19.53	3.29	0.46	2.62	19.38
Multi-Turn	LSTM w/ attn	7.06	2.34	3.18	22.76	4.74	1.10	3.36	19.63
	Transformer	7.35	2.65	4.06	25.91	4.64	1.21	3.53	16.75
	T5-small	9.49	3.81	4.77	25.83	7.38	2.42	2.81	9.77
	GPT-2	8.55	3.39	5.12	27.75	7.26	2.28	3.13	12.24
	AdaLabel	6.13	2.11	4.63	28.65	5.75	1.41	3.71	14.77
	DialogBERT†	6.34	1.88	5.21	30.61	3.90	0.68	3.03	22.01

Table 4: Performance of various models on DailyDialog and OpenSubtitles. †DialogBERT uses sampling-based decoding following the original implementation. Other models use greedy decoding; we observe DialogBERT failed to perform reasonably with greedy decoding.

tributed utterance order ranking, so as to better capture discourse-level dialogue context.

In our experiments, we specifically choose AdaLabel and DialogBERT because they are recently published models from top-tier venues using overlapping datasets. Moreover, the authors publish their code[7], allowing us to replicate their work on our cleaned datasets.

Setup. We use the fairseq (Ott et al., 2019) framework to train the LSTM model and the Transformer model. The LSTM model has a single layer with a hidden size of 512 for both the encoder and decoder. The Transformer model has 6 layers with a hidden size of 512 with 8 attention heads. Our pretrained models are adopted from HuggingFace (Wolf et al., 2019) and fine-tuned in the experiments. Specifically, T5-small uses 6 layers with a hidden size of 512 and 8 attention heads, whereas GPT-2 uses 12 layers with a hidden size of 768 and 12 attention heads. For AdaLabel and DialogBERT, we directly run their source code on our cleaned datasets. For AdaLabel, both the encoder and decoder contain 6 layers with a hidden size of 512 and 8 attention heads. DialogBERT uses 6 transformer layers, a hidden size of 256, and 2 attention heads for the utterance encoder, context encoder, and decoder. All models are trained with the Adam optimizer (Kingma and Ba, 2015).

5.3. Results

Table 4 shows the performance of baseline and state-of-the-art models on our cleaned datasets. For completeness, we evaluate all models on both single-turn and multi-turn settings, which is common in previous studies (Sun et al., 2021; Cai et al., 2020b; Wang et al., 2021).

As seen, the simple LSTM model is on par with the Transformer model: it is slightly worse on DailyDialog but better on OpenSubtitles. The pretrained models outperform the un-pretrained ones, which is consistent with existing literature such as machine translation (Lewis et al., 2020; Liu et al., 2020; Raffel et al., 2020).

However, the performance gap between different baseline models is noticeably smaller in other text generation tasks. For example, Luong et al. (2015) achieve a BLEU score of 20.9 on the WMT'14 English–German dataset using LSTM with attention, whereas Vaswani et al. (2017) achieve 27.3 with a Transformer model, giving an improvement of 6.4 points. Raffel et al. (2020) further improve the model by 3.6 points with pretraining techniques, achieving a BLEU score of 30.9. We hypothesize that this is due to the inherent uncertainty of the dialogue task, which may not be fully alleviated by pretraining techniques.

We then compare the above baselines with several alleged state-of-the-art models: AdaLabel (Wang et al., 2021) and DialogBERT (Gu et al., 2021), because their models are open-sourced.

AdaLabel is a Transformer-based encoder–decoder model, enhanced with an adaptive label smoothing technique. Based on the cleaned datasets, we find that AdaLabel's performance is (at most) on par with the pretrained models. On the one hand, AdaLabel yields consistently lower BLEU scores than GPT-2. On the other hand, AdaLabel appears to achieve higher diversity scores in some settings, but is much lower in others (e.g., single-turn DailyDialog).

Another alleged state-of-the-art model is DialogBERT, which uses hierarchical BERT to encode multi-turn context. In our experiments, we were unable to obtain reasonable performance with greedy decoding since the model soon converges to generic responses. Therefore, we keep the the same sampling setting as in the original paper, which leads to high diversity scores

[7]AdaLabel: `https://github.com/lemon234071/AdaLabel`; DialogBERT: `https://github.com/guxd/DialogBERT`

(given by the Dist metric). However, a higher diversity score does not necessarily imply a better dialogue system, as a random decoder will achieve the highest Dist scores. In fact, DialogBERT results in very low BLEU scores (the main metric in most literature): it is worse than all models in all settings, except the LSTM in single-turn DailyDialog.

Overall, it is highly questionable whether these alleged state-of-the-art models truly outperform standard baselines. Our observations contradict the original papers' experiments (Wang et al., 2021; Gu et al., 2021), where the authors use overlapping datasets and show the superior performance of the proposed models in terms of both BLEU and Dist. Such a discrepancy is largely due to the overlapping samples: although their proposed models achieve a lower BLEU score in terms of conversational skills, they can boost it by memorizing overlapping samples, showing unreal improvement in both BLEU and Dist. This confirms the importance of cleaning existing benchmark datasets for future dialogue research.

6. Conclusion

In this work, we observe the overlapping issues in two popular open-domain dialogue benchmark datasets: DailyDialog and OpenSubtitles. We then conduct systematic analyses and show the undesired consequences when data overlap. To address the issue, we propose a data cleaning strategy to set up a proper protocol for future research. Our experiments on the cleaned datasets show that the "state-of-the-art" performance on overlapping datasets is questionable, highlighting the importance of revisiting open-domain dialogue datasets.

7. Acknowledgments

The research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant No. RGPIN2020-04465, the Amii Fellow Program, the Canada CIFAR AI Chair Program, a UAHJIC project, a donation from DeepMind, and Compute Canada (www.computecanada.ca).

8. Bibliographical References

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Bahuleyan, H., Mou, L., Zhou, H., and Vechtomova, O. (2019). Stochastic wasserstein autoencoder for probabilistic sentence generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4068–4076.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gasic, M. (2018). MultiWOZ – A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Cai, H., Chen, H., Song, Y., Zhang, C., Zhao, X., and Yin, D. (2020a). Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343.

Cai, H., Chen, H., Zhang, C., Song, Y., Zhao, X., Li, Y., Duan, D., and Yin, D. (2020b). Learning from easy to complex: Adaptive multi-curricula learning for neural dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7472–7479.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186.

Dong, L. and Lapata, M. (2016). Language to logical form with neural attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 33–43.

Du, J., Li, W., He, Y., Xu, R., Bing, L., and Wang, X. (2018). Variational autoregressive decoder for neural response generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3154–3163.

El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. (2017). Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Gu, X., Yoo, K. M., and Ha, J.-W. (2021). DialogBERT: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12911–12919.

Guo, J., Liu, Q., Lou, J.-G., Li, Z., Liu, X., Xie, T., and Liu, T. (2020). Benchmarking meaning representations in neural semantic parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1520–1540.

Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Huang, C., Yang, W., Cao, Y., Zaïane, O., and Mou, L. (2021). A globally normalized neural model for semantic parsing. In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 61–66.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1):26–41.

Khan, K., Sahu, G., Balasubramanian, V., Mou, L., and Vechtomova, O. (2020). Adversarial learning on the latent space for diverse dialog generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5026–5034.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the International Joint Conference on Natural Language Processing*, volume 1, pages 986–995.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1742–1748.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., and Jin, Z. (2016). Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–53.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Schumann, R., Mou, L., Lu, Y., Vechtomova, O., and Markert, K. (2020). Discrete optimization for unsupervised sentence summarization with word-level extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5032–5042.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3776–3783.

Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, volume 1, pages 1577–1586.

Sun, B., Feng, S., Li, Y., Liu, J., and Li, K. (2021). Generating relevant and coherent dialogue responses using self-separated conditional variational autoencoders. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, volume 1, pages 5624–5637.

Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., and Zhao, D. (2017). How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wallace, R. S. (2009). The anatomy of A.L.I.C.E. In *Parsing the Turing Test*, pages 181–210.

Wang, Y., Zheng, Y., Jiang, Y., and Huang, M. (2021). Diversifying dialog generation via adaptive label smoothing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, volume 1, pages 3507–3520.

Weizenbaum, J. (1966). ELIZA–A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9(1):36–45.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gasic, M., Barahona, L. M. R., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 438–449.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xu, Z., Liu, B., Wang, B., Sun, C., Wang, X., Wang, Z., and Qi, C. (2017). Neural response generation via GAN with an approximate embedding layer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 617–626.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2204–2213.

Zhou, L., Gao, J., Li, D., and Shum, H.-Y. (2020). The design and implementation of Xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Zhou, W., Li, Q., and Li, C. (2021). Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, volume 1, pages 694–703.

Zhu, Q., Huang, K., Zhang, Z., Zhu, X., and Huang, M. (2020). CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.