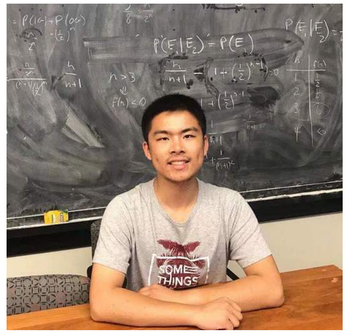
An Empirical Study on the Overlapping Problem of Open-Domain Dialogue Datasets

Yuqiao Wen Guoqing Luo Lili Mou


yuqiao@ualberta.ca gluo@ualberta.ca
doublepower.mou@gmail.com

Dept. Computing Science, University of Alberta
Alberta Machine Intelligence Institute (Amii)

LREC-2022

Outline

- **Introduction**
- Bizarre Behaviours
- Dataset Cleaning
- Model Performance
- Conclusion

Dialogue Generation

- Task-Oriented
- Open-Domain
  - Open-ended conversations
  - Application: chatbot

Example:

A: Do you believe in horoscope fortune-telling?

B: I used to be an atheist, but in recent months, I couldn’t but form a more favourable opinion of horoscope.

Dialogue Datasets

- DailyDialog
  - daily life conversations
- OpenSubtitles
  - dialogues from movies
- Widely used for dialogue research

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*, 2017.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC,* 2018.

The Overlapping Problem

- (Near-)Identical samples between the
  - training set
  - test set
- Consequences
  - Inflated performance
  - Arbitrary performance
  - Over-informative output

DailyDialog: ~23%
OpenSubtitles: ~34%

Our Contributions

- Addressing the overlapping problem
- Performing systematic analyses
- Proposing a data cleaning strategy

Outline

- Introduction
- **Bizarre Behaviours**
- Dataset Cleaning
- Results
- Conclusion

Overlap Statistics

- $\mathbf{u} = \{u_1, \cdots, u_m\}, \mathbf{v} = \{v_1, \cdots, v_n\}$
- $R(\mathbf{u}, \mathbf{v}) = \dfrac{2\,|\mathbf{u} \cap \mathbf{v}|}{|\mathbf{u}| + |\mathbf{v}|}$
- sample: $\mathbf{x} = (\mathbf{c}, \mathbf{r})$
- $R(\mathbf{x}, \mathbf{x}') = \min\{R(\mathbf{c}, \mathbf{c}'), R(\mathbf{r}, \mathbf{r}')\}$
- $R(\mathbf{x}, \mathscr{D}_{\text{train}}) = \max_{\mathbf{x}' \in \mathscr{D}_{\text{train}}} R(\mathbf{x}, \mathbf{x}')$

Overlap Statistics

- Significant Overlapping

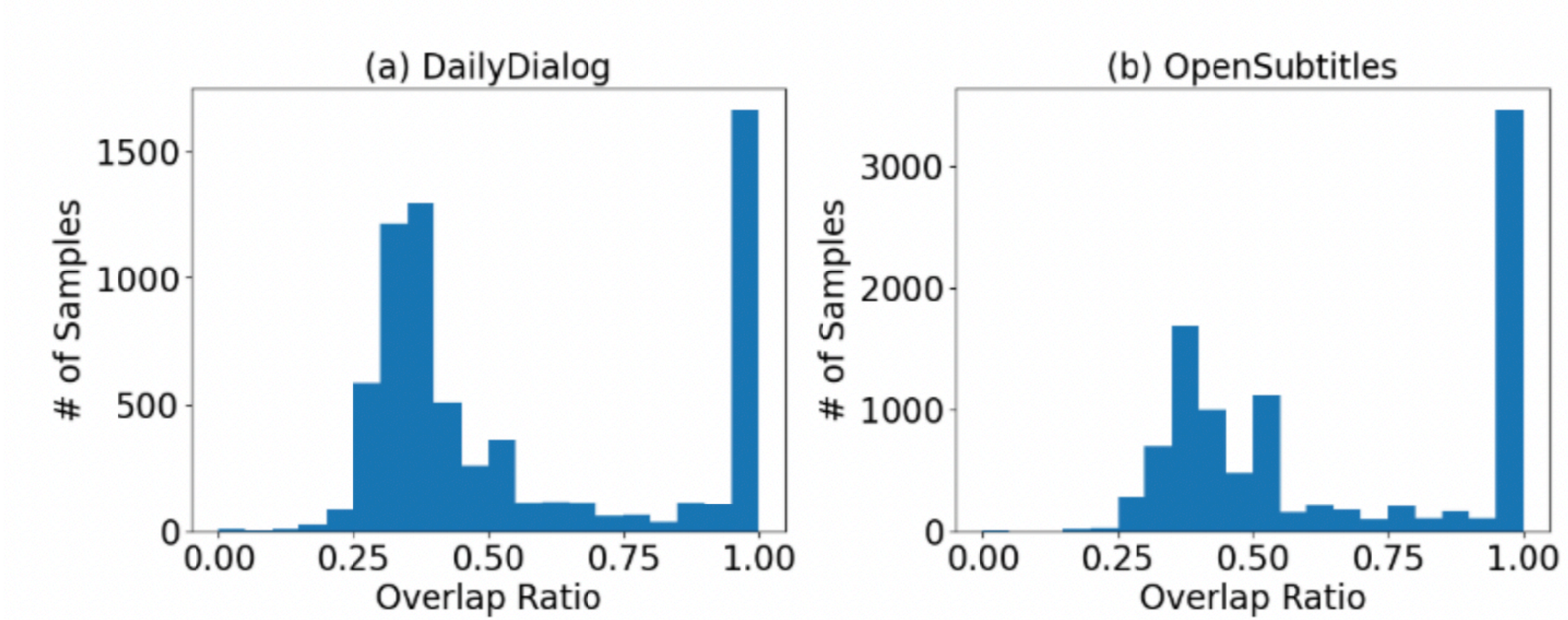


Figure 1: Overlap histogram of test samples against the training set on (a) DailyDialog and (b) OpenSubtitles

Overlap Statistics

- Significant Overlapping

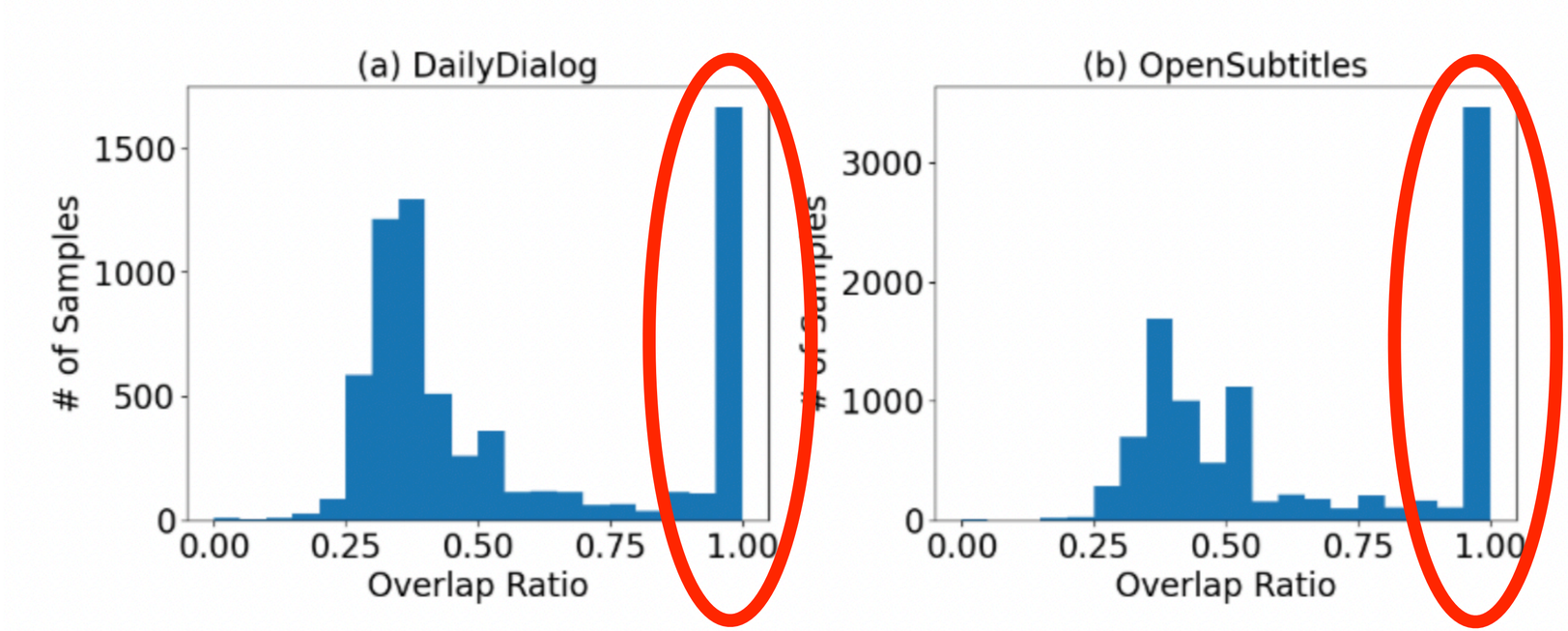


Figure 1: Overlap histogram of test samples against the training set on (a) DailyDialog and (b) OpenSubtitles

Examples

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

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

Overlap Statistics

- Significant Overlapping

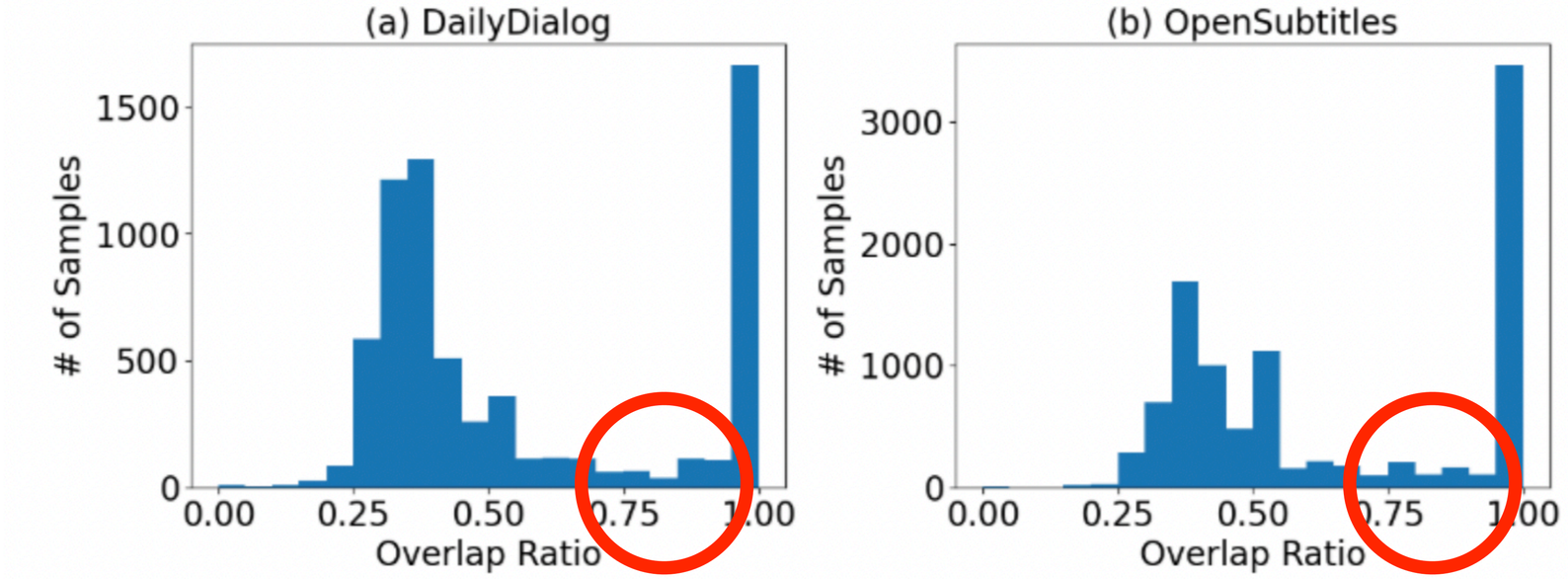


Figure 1: Overlap histogram of test samples against the training set on (a) DailyDialog and (b) OpenSubtitles

Examples

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

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

Overlap Statistics

- Significant Overlapping

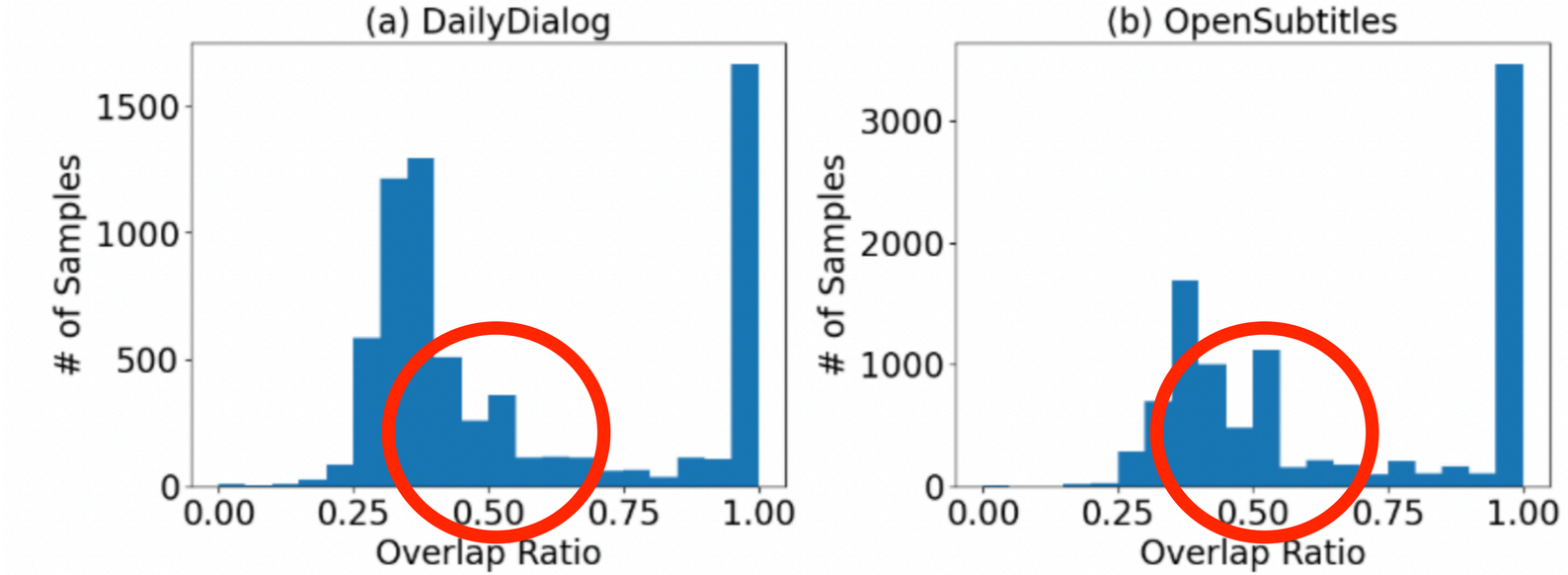


Figure 1: Overlap histogram of test samples against the training set on (a) DailyDialog and (b) OpenSubtitles

Examples

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.

DailyDialog

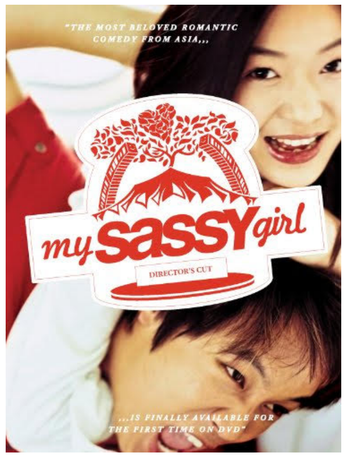
- Dataset Construction
  - crawled from English learning websites
- Potential Causes
  - similar learning materials

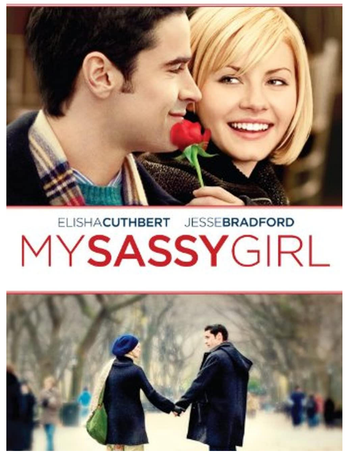
OpenSubtitles

- Dataset Construction
  - extracted from subtitle files
  - organized by IMDb identifiers
- Cause of Overlapping
  - remakes of the same movies

Example

- My Sassy Girl
  - Original Release (2001)
  - American Remake (2008)





- Different IMDb identifiers
- Highly overlapping dialogues

Bizarre Behaviours

- Inflated Performance
- Arbitrary Performance
- Over-informative Responses

Inflated Performance

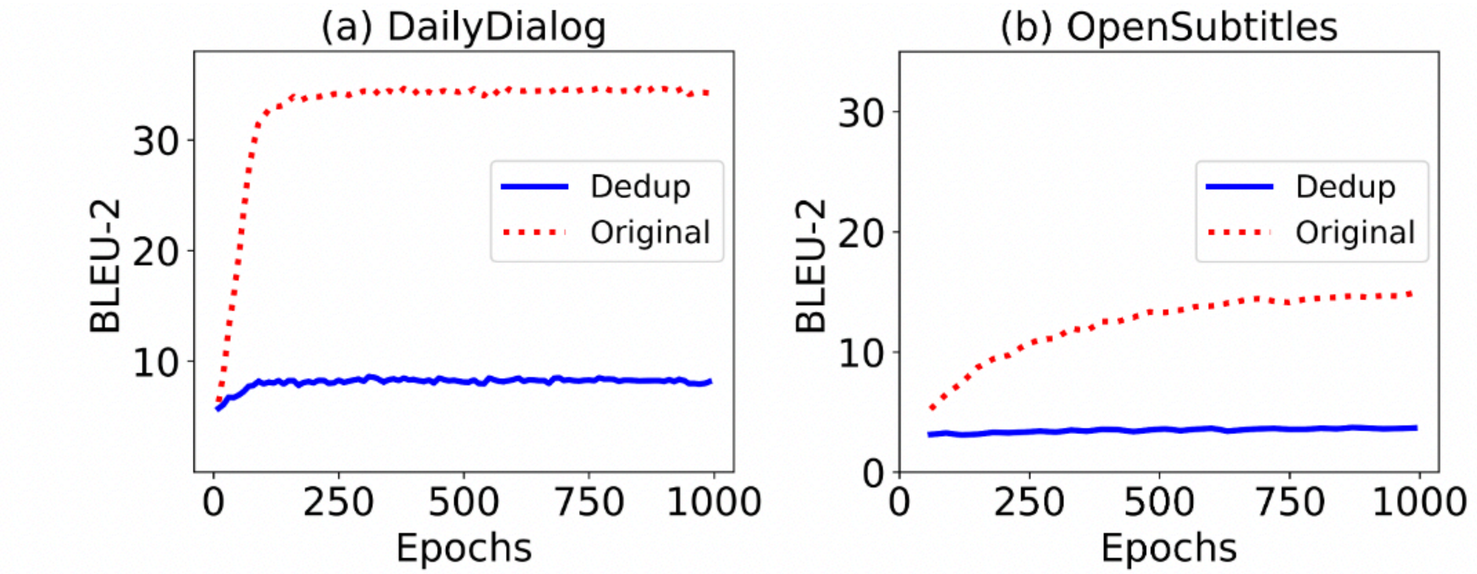


Figure 2: BLEU-2 learning curve comparison between the original dataset and the deduplicated dataset for (a) DailyDialog and (b) OpenSubtitles. Samples with an overlap greater than 0.80 are considered duplicates and are removed for the deduplicated dataset.

Arbitrary Performance

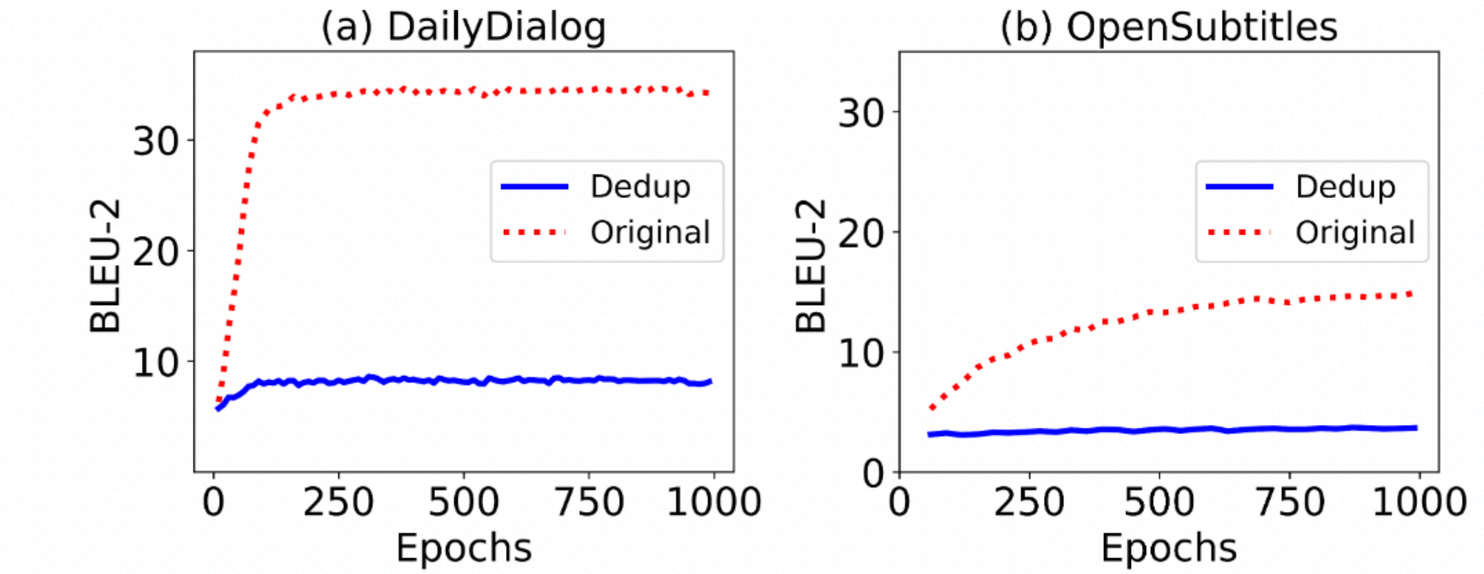


Figure 2: BLEU-2 learning curve comparison between the original dataset and the deduplicated dataset for (a) DailyDialog and (b) OpenSubtitles. Samples with an overlap greater than 0.80 are considered duplicates and are removed for the deduplicated dataset.

Over-informative Output

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 .

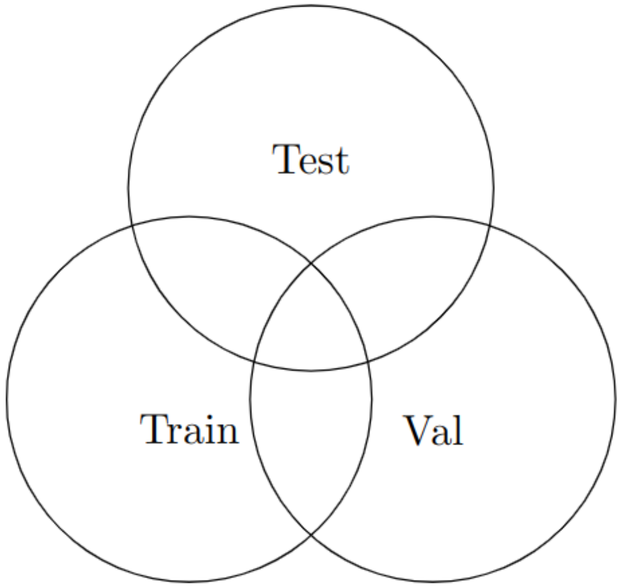
OpenSubtitles	
Train Input	But you have some strength in you , my dear Hobbit .
Train Ref	What happened, Gandalf ?
Test Input	But you have some strength in you , my dear Hobbit .
Test Ref	What happened, Gandalf ?
Model Output	What happened, Gandalf ?

Outline

- Introduction
- Bizarre Behaviours
- **Dataset Cleaning**
- Results
- Conclusion

Deduplication

- Deduplicate the original sets
  - at least one of validation/test will shrink
  - unreliable validation/test performance
- Deduplicate all, then re-split
  - allows us to keep large validation/test sets



Deduplication

- Deduplicate and re-split in multi-turn setting
  - DailyDialog: dialogue session
  - OpenSubtitles: entire movie
- Avoid information leaking
- Accommodate both multi-turn and single-turn setting

Overlap Ratios

- $\mathscr{D}$: set of all samples
- $\mathbf{u}$, $\mathbf{v}$: two samples
- $$R(\mathbf{u}, \mathbf{v}) = \frac{2\,|\mathbf{u} \cap \mathbf{v}|}{|\mathbf{u}| + |\mathbf{v}|}$$
- $$R(\mathbf{u}, \mathscr{D}) = \max_{\mathbf{u}' \in \mathscr{D} \setminus \{\mathbf{u}\}} R(\mathbf{u}', \mathbf{u})$$

Overlap Ratios

- $\mathscr{D}$: set of all samples
- $\mathbf{u}$, $\mathbf{v}$: two samples
- $$R(\mathbf{u}, \mathbf{v}) = \frac{2\,|\,\mathbf{u} \cap \mathbf{v}\,|}{|\,\mathbf{u}\,| + |\,\mathbf{v}\,|}$$
- $$R(\mathbf{u}, \mathscr{D}) = \max_{\mathbf{u}' \in \mathscr{D} \setminus \{\mathbf{u}\}} R(\mathbf{u}', \mathbf{u})$$

Before

- $\mathbf{u} = \{u_1, \cdots, u_m\}$, $\mathbf{v} = \{v_1, \cdots, v_n\}$
- $$R(\mathbf{u}, \mathbf{v}) = \frac{2\,|\,\mathbf{u} \cap \mathbf{v}\,|}{|\,\mathbf{u}\,| + |\,\mathbf{v}\,|}$$
- sample: $\mathbf{x} = (\mathbf{c}, \mathbf{r})$
- $R(\mathbf{x}, \mathbf{x}') = \min\{R(\mathbf{c}, \mathbf{c}'), R(\mathbf{r}, \mathbf{r}')\}$
- $$R(\mathbf{x}, \mathscr{D}_{\text{train}}) = \max_{\mathbf{x}' \in \mathscr{D}_{\text{train}}} R(\mathbf{x}, \mathbf{x}')$$

Deduplication

- Compute overlap ratios
- Remove overlapping samples
- Repeat until clean

Outline

- Introduction
- Bizarre Behaviours
- Dataset Cleaning
- **Results**
- Conclusion

Models

- Standard models
  - LSTM w/ attention
  - Transformer
  - T5-small
  - GPT-2
- “State of the art”
  - AdaLabel
  - DialogBERT

Yida Wang, Yinhe Zheng, Yong Jiang, Minlie Huang. Diversifying Dialog Generation via Adaptive Label Smoothing. In *IJCNLP*, 2021.

Xiaodong Gu, Kang Min Yoo, Jung-Woo Ha. DialogBERT: discourse-aware response generation via learning to recover and rank utterances. In *AAAI*, 2021

Model Performance

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

Model Performance

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

Model Performance

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

AdaLabel: training with smoothed labels

DialogBERT: contextual modeling w/ hierarchical BERT

Conclusion

- Observe the overlapping problem
- Perform systematic analysis
- Provide cleaned datasets

Take-home Messages

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

Acknowledgements

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).

References

Alec Radford, Jefferey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. Language models are unsupervised multitask learners. In *EMNLP*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need. In *NIPS*, 2017

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text Transformer. In JMLR, 2020.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC,* 2018.

Xiaodong Gu, Kang Min Yoo, Jung-Woo Ha. DialogBERT: discourse-aware response generation via learning to recover and rank utterances. In *AAAI*, 2021

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*, 2017.

Yequan Wang, Minlie Huang, Li Zhao, Xiaoyan Zhu. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*, 2016.

Yida Wang, Yinhe Zheng, Yong Jiang, Minlie Huang. Diversifying Dialog Generation via Adaptive Label Smoothing. In *IJCNLP*, 2021.

Questions?

yuqiao@ualberta.ca

Thank you!