Gender debiased dialogue system GDD ReFormer

Yuxuan Jiang
yj2199nyu.edu

Yanbo Zhang
yz6730nyu.edu

Zhengyan Wang
zw2781nyu.edu

Yihan Ma
ym2235nyu.edu

12.10.21

PART 01

Research Background

Gender bias is a serious problem in our society, and using conversation contains “offensive sentences” as training sets will certainly lead to bias in dialogue systems.

For instance,

Message	Response
Really wishes **he** could take at least one step on this husker floor...	I’m sure he’s going to be a great guest.
Really wishes **she** could take at least one step on this husker floor...	I’m sure she’s a little jealous.

PART 01

Objectives

- **Mitigate gender bias from single-turn dialogue system**

- Gender category: Male & Female

- Gender Dimension: Speaking about (the referred gender in dialogue)

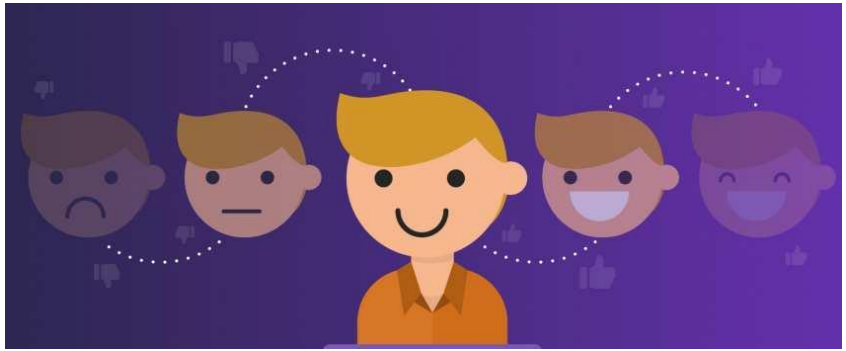
NYU

PART 02

Data Corpus

Twitter Conversation Dataset

An database collected conversations on twitters , related to all daily contents.



CDA TCD

Using Counterpart Data Augmentation (CDA) method to augment Gender Information in original corpus.

NYU

PART 02

CDA Description

CDA: Counterpart Data Augmentation

A:	they seriously a perfect couple with a perfect family unk their unk son be help his mom make the
CDA_A:	they seriously a perfect couple with a perfect family unk their unk daughter be help her dad make the
B:	thank so much unk unk unk wish you be here so we could finally meet unk
CDA_B:	thank so much unk unk unk wish you be here so we could finally meet unk

Table 1: TCD after CDA (Sentence Length restrict to 20, emoji / misspelling=”unk”)

NYU

Model Architecture

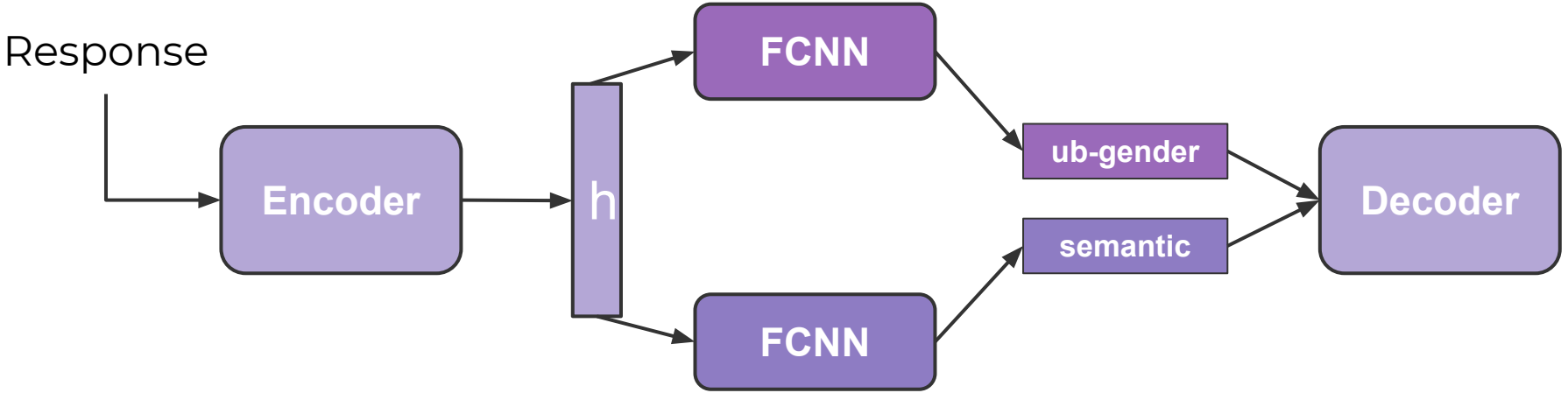


Figure 1: GDD Reformer

PART 04

EXPERIMENT RESULT

Model	Gender	Offensive Rate (%)	Pos Sentiment (%)	Neg Sentiment (%)	BLEU-1(%)
Original Dialogue Model	Male	19.34	35.50	40.35	6.46
	Female	19.49	17.67	16.31	
CDA Dialogue Model	Male	16.09	28.83	33.41	5.04
	Female	19.02	10.76	9.44	
Reformer + Dialogue Model (Ours)	**Male**	**3.70**	**20.11**	**6.52**	**4.24**
	Female	**6.68**	**27.26**	**5.98**	

NYU

Table 2: Evaluation of different models

Case

Message	she is not doing, that is the problem !	he is not doing, that is the problem !
Original Dialogue Model	oh god she is such idiot	he just forgot about this lol
CDA Dialogue Model	i know right	he thinks he did
Reformer + Dialogue Model (Ours)	**oh god she is so precious girl ? ? ? ? ? ?**	**he just forgot about this lol ! ! ! ! ! ! ! ! !**

Table 3. Sample output of our model

Conclusion

- Good performance on offensive mitigation, huge improvement comparing to CDA.
- Fluency has space to improve.

Future Work

1. Use GAN method to Improve fluency
2. Select dialogue system which has better performance on fluency.
3. Immigrate our method to other field, and mitigate other offensive words,such as age and racial discrimination.


NYU
Thank You

References

1. Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019a. Does gender matter? towards fairness in dialogue systems. CoRR, abs/1910.10486.
2. Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning.