IntentCONANv2 Intent-Specific Counterspeech Generation

SOMESH AGRAWAL(2024AIB2292)

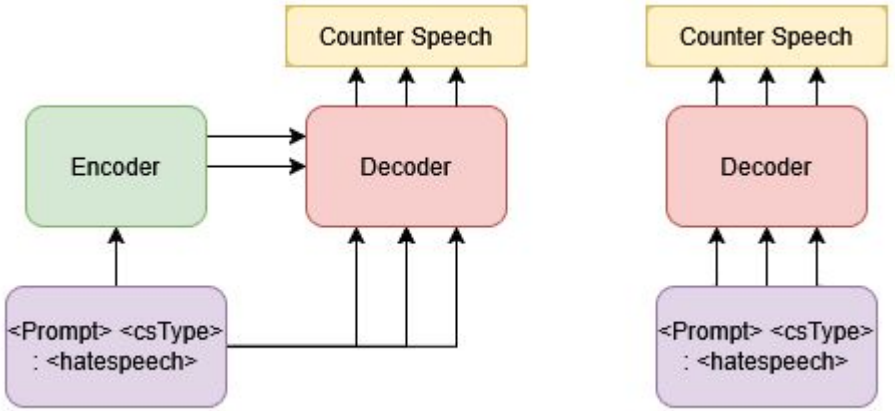
Overview

All Experiments Performed
Limitations
Future Works

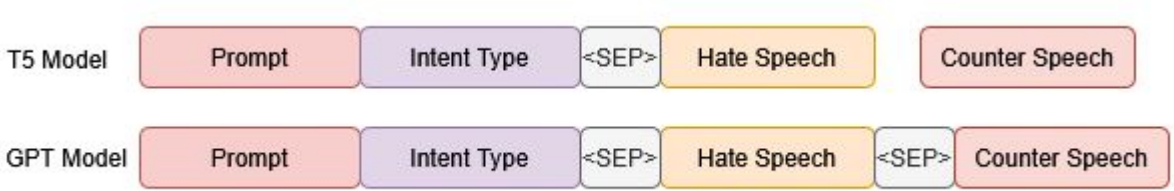
DATASET

HateSpeech->	Juice Jews are worse that nukes. Caption to a meme showing Hiroshima 1945 and present day, contrast against pictures of Detroit in 1945 and present day.
Informative	Anti-semitism is a serious problem that we need to address and stop spreading relentless hate against people who are innocent.
Questioning	Want to try a nuke then?
Denouncing	Really? i'd sit with a group of jews any day.
Positive	We need to understand from where this hate stems from. Stop hating, start loving. The human race needs to be more appreciative of one another.

Experiment-1



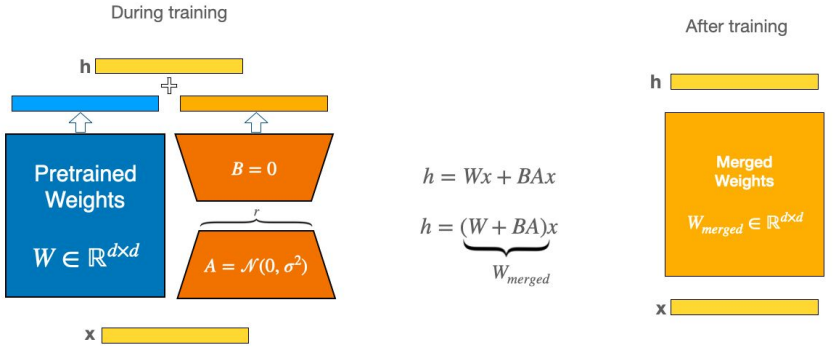
Difference in Input Prompt for Encoder-Decoder and Decoder Only Model

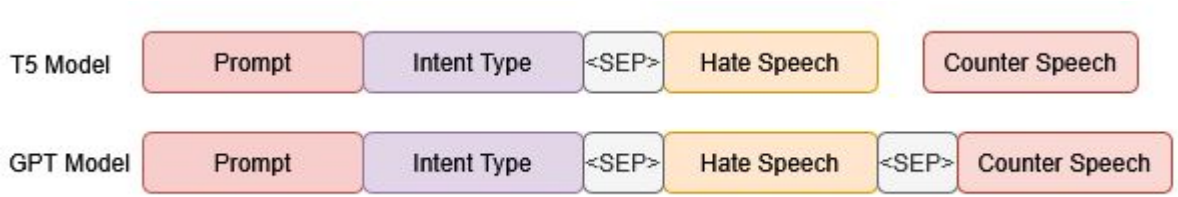


Prompts Used for Different Models

- Used GPT2 and T5 Model for End to End Finetuning
- Prompts Used : Generate <cs_type>: hateSpeech
- T5 generally produced more coherent and intent-aligned counterspeech with this basic prompting strategy compared to GPT-2.(Due to it's bidirectional feature learning)
- T5-Base was used as the final structure for future model updates.
- Finetuning was taking time per epochs.

Experiment-2: PEFt using LORA





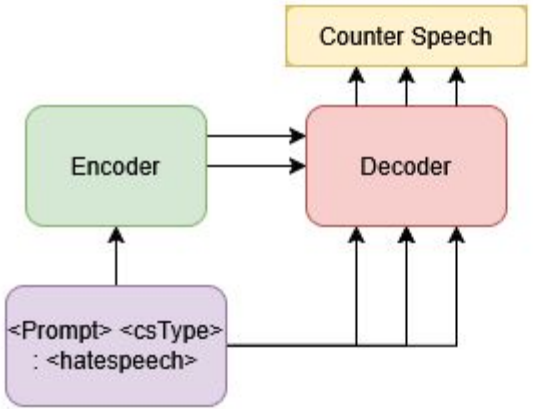
Prompts Used for Different Models

- Use t5-base with Lora Adapters for finetuning.
- Main Aim is to get a faster model(less computational time)with minimal loss on accuracy and metric scores (BLEU,BERT, ROUGE).

Result:

- Model trained faster, but the loss was very high w.r.t. metrics on finetuned model.
- Full Finetuning is better irrespective of time complexity.

Experiment-3- Two-Stage General and Intent-Specific Fine-tuning



T5-Encoder Decoder Model

- Stage Fine-tuning strategy was employed: t5-base was initially adapted to the counterspeech domain using all data (without intent conditioning), (on train data)
- followed by a second stage of fine-tuning focused on intent-specific generation using structured prompts. (train+validation data)
- **Reason**: Initial stage builds general counterspeech knowledge. The second stage refines it based on intent-specific structure.

Result:

- Model overfitted to the training data.
- Suggests that direct intent-conditioned training is more effective for this task.
- **Full fine-tuning** remains more robust, despite higher computational cost.

Experiment-4&5- Prompt Augmentation

Original Prompt: generate <cs_type>: <hate speech>

Additional Templates:(Method 4)
“generate a counter speech with intent {cs_type} : {hatespeech} ”
generate [<cs_type>]: [<hate speech>]
write a response reflection {cs_type} to : {hatespeech}

Add noise Inputs(Method 5)
generate <cs_type>: <hate speech> ???? :::”” produce <cs_type>: <hate speech>
yield <cs_type>: <hate speech>

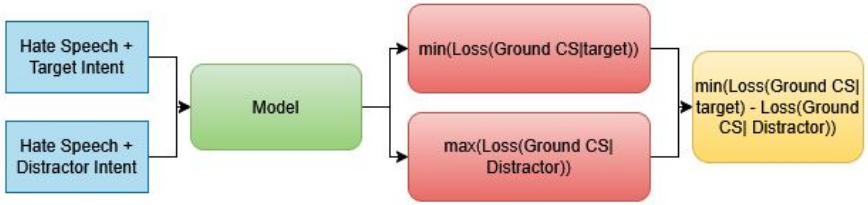
Issue: Model was sensitive to minor changes in prompt to produce the final generated sentence, even tough Intent is kept same.

- **To improve robustness to prompt phrasing, Prompt Augmentation was implemented.**
- **This included training with (Method 4) a predefined set of templates instructional phrasings and**
- **(Method 5) dynamic perturbations like typo introduction and synonym replacement within the prompt's instructional component.**

Result:

- Techniques like prompt augmentation provided modest improvements by enhancing data diversity and model robustness.

Experiment-6(Novelty):- Intent Contrastive Learning



(Distractor Forward Pass):

$$L_{distractor} = NLL(Model(Prompt_{distractor}, CS_{gold}))$$

Inverted Loss:

$$L_{away} = -\mathcal{L}_{distractor}$$

Combined Loss:

$$L_{combined} = \mathcal{L}_{target} + \beta \cdot \mathcal{L}_{away}$$

Optimization Step:

Backward and update on $L_{combined}$

Main Features:

- Rather than training a DPO, PPO or any other RL Model, we can use the same model to learn to differentiate between actual or intended output.
- During training, the model was simultaneously fed a positive example (correct hate speech-intent pairing) and a negative example (same hate speech, deliberately mismatched intent), both with instruction prompts.
- A composite loss function, combining standard generation loss with a contrastive term, penalized alignment with the incorrect intent and rewarded clear differentiation, fostering robust intent adherence.
- Composite loss conveys the Loss over the Ground Counter Speech given the sentence generated from input (intent target and hate speech). Both losses are considered, and we try to mnimze the loss w.r.t to sentence generated by target intent and maximize wrt to output generated by Distractor intent.

Results

Table 1: Evaluation metrics across different experiments

c

Experiment	Loss	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore (F1)
Full Fine-tuned T5	0.303	21.804	6.248	16.347	12.454	86.718
Fully Finetuned GPT2	1.819	16.545	4.322	10.432	9.334	80.340
Fine-tuned T5 with LoRA Adapters	0.392	16.709	3.434	12.714	10.930	86.395
Two-Stage Fine Tuning	0.308	17.607	5.432	15.619	**15.143**	87.103
Fixed Template Prompt Augmentation	0.298	22.471	**7.818**	**18.432**	14.320	87.130
Dynamic Prompt Augmentation	**0.293**	**22.521**	7.192	18.194	13.650	86.239
Intent Contrastive Self-Correction (ICS)	0.301	22.450	6.385	17.653	13.445	**88.143**

Thanks!!