IntentCONANv2 Intent-Specific Counterspeech Generation

Somesh Agrawal(2024AIB2292)[1]

[1]Indian Institute of Technology Delhi

May 10, 2025

Abstract

Mitigating online hate speech is crucial, and intent-specific counterspeech offers a promising, nuanced approach. This research focuses on developing models for this task using the IntentCONANv2 dataset, which comprises 13K hate speech-counterspeech pairs conditioned on four intents: informative, denouncing, question, and positive. We investigate the fine-tuning of two pre-trained Transformer architectures, T5 (a sequence-to-sequence model) and GPT-2 (a decoder-only model), to generate counterspeech conditioned on input hate speech and a specified csType (intent). Model performance is evaluated using BLEU, ROUGE, and BERTScore.

Keywords
T5, GPT3, Finetuning, PEFT
Github
https://github.com/somesh2002/counter-speech-intent-conan-v2

1 Introduction

The rise of online hate speech necessitates effective countermeasures beyond simple content removal. Counterspeech, which directly challenges or refutes hateful content, offers a proactive means to educate, support victims, and influence social norms (Benesch et al., 2016). The impact of counterspeech can be significantly amplified when its communicative intent is carefully chosen (Chung et al., 2021). This project addresses the challenge of generating such intent-specific counterspeech

The IntentCONANv2 dataset (Gomez et al., 2023) provides a valuable resource for developing intent-aware counterspeech generation systems. This project aims to leverage this dataset to build a model capable of generating high-quality speech that is not only relevant to the input but also accurately reflects the desired communicative intent.

Our primary objective is to fine-tune a pre-trained language model to perform this conditional generation task. We hypothesize that by explicitly providing the target intent as part of the input, the model can learn to generate more diverse and targeted counterspeech.

2 Methodology

2.1 Dataset

This research leverages the IntentCONANv2 dataset, a resource containing approximately 13,000 entries specifically designed for intent-conditioned counterspeech generation(at 80-15-5 split for train, validation and test). Each entry comprises a piece of hate speech, a corresponding human-written counterspeech, and a crucial csType label. This label indicates the primary intent of the counterspeech, falling into one of four categories: informative (providing facts or evidence), denouncing (condemning the hate speech), question (posing a query to challenge or provoke thought), or positive (offering support or promoting positive values).

2.2 Data Preprocessing

We need to prepare the IntentCONANv2 dataset for model training and validation. For each instance, the input to the model was formed by concatenating a special token representing the target intent. The model's target output during training was the corresponding counterspeech text. All texts were then tokenized using the model tokenizer, and necessary padding and truncation was applied.

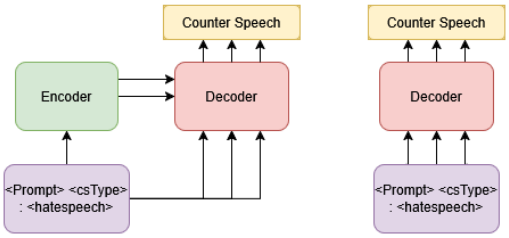

Figure 1: Prompt Inputs to Encoder-Decoder(Left) vs Decoder Only(Right) Model on Training

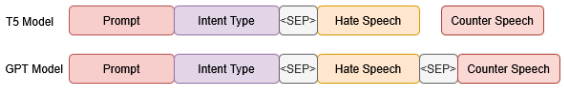

Figure 2: Different prompt inputs based on model architecture. In T5, the gold counterspeech is provided as a label, while for GPT, it is appended to the input during training. Prompt variation can be introduced via modified initial tokens, added noise, or diverse template formats.

2.3 Model Architectures and Training Strategies

We explored two main Transformer-based architectures: T5 (Raffel et al., 2020), a versatile text-to-text encoder-decoder model, and GPT-2 (Radford et al., 2019), a powerful decoder-only language model.

We began by conducting an initial exploratory phase using both pre-trained t5-base and gpt2 models from the Hugging Face Model Hub. In this stage, we performed direct inference by constructing prompts combining the target counterspeech type (csType) and the original hate speech.

Qualitative assessment of the generated outputs revealed that T5 generally produced more coherent and intent-aligned counterspeech with this basic prompting strategy compared to GPT-2. Given these initial observations and T5's inherent suitability for conditional text generation tasks, we selected t5-base as the primary architecture for subsequent, more intensive fine-tuning and experimentation. Below are the several experiments done by us.

Parameter Efficient FineTuning (PEFT) : Employed Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) on the T5-base model to facilitate longer training with reduced computational overhead by updating only a small subset of the model's parameters. This approach enhances resource efficiency and accelerates model inference through adapter-based modifications, enabling the execution of more complex experiments.

Two-Stage General and Intent-Specific Fine-tuning : A Two-Stage Fine-tuning strategy was employed: t5-base was initially adapted to the counterspeech domain using all data (without intent conditioning), followed by a second stage of fine-tuning focused on intent-specific generation using structured prompts.

Prompt Augmentation : 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.

Intent Contrastive Self-Correction(ICS) : It was introduced to significantly enhance intent discrimination. 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 augmented 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.

2.4 Training Stratergy

The t5-base) model was fine-tuned on the preprocessed training set using the Hugging Face Transformers. We employed the **AdamW optimizer** with a learning rate of 1e-4 and a batch size of 16. Training typically ran for 5-10 epochs, with early stopping based on validation loss to prevent overfitting, using standard cross-entropy loss as the objective function. For generating counterspeech during inference, beam search (beam size of 5) was utilized, with a maximum output length of 256 tokens.

3 Results

3.1 Evaluation Metrics

The quality of the generated counterspeech was evaluated using the following standard metrics:

- **BLEU** (Papineni et al., 2002): Measures n-gram precision between generated and reference texts.
- **ROUGE** (Lin, 2004): Measures n-gram recall (ROUGE-N) and longest common subsequence overlap (ROUGE-L). We report ROUGE-1, ROUGE-2, and ROUGE-L scores.

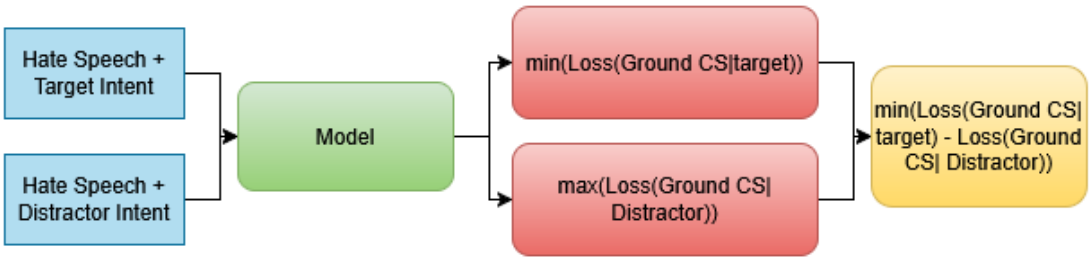

Figure 3: During training, both the true intent and a distracted (incorrect) intent are provided. The model is optimized to minimize the loss with respect to the true intent while simultaneously maximizing the loss with respect to the distracted intent, using the ground truth counterspeech as the target after the model generation.

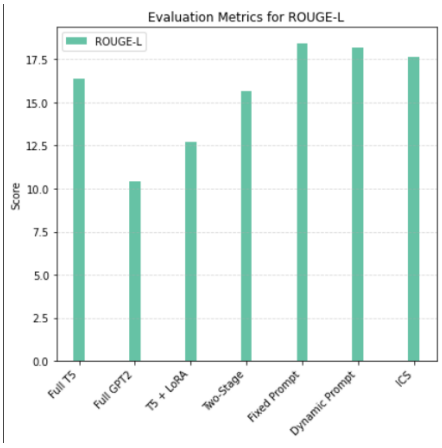

Figure 4: Box Plot for the Rouge-L Score for Different Experiments.

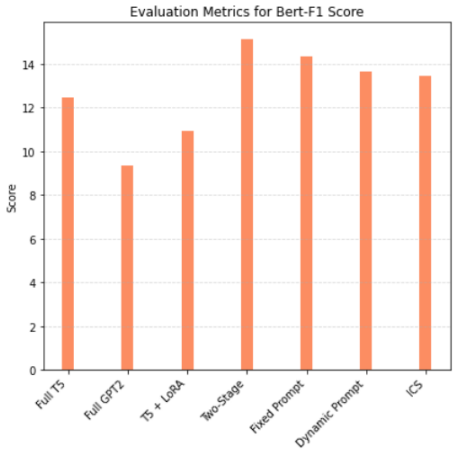

Figure 5: Box Plot for the Bleu Score for Different Experiments.

- **BERTFScore** (Zhang et al., 2019): Evaluates semantic similarity using contextual BERT embeddings. We report the BERTScore F1.

These metrics offer complementary perspectives: BLEU and ROUGE assess lexical similarity, while BERTScore captures semantic alignment—essential for evaluating the nuanced intent and meaning in counterspeech.

3.2 Quantitative Results

The differnt models and architectures was evaluated on the held-out test set. The results, aggregated overall and broken down by intent, are presented in Table 1. And Plots have been given in Figure 4 & 5.

4 Discussion

Our experiments revealed several critical factors influencing intent-specific counterspeech generation. The superior performance of BERTScore over n-gram metrics like BLEU and ROUGE highlights the importance of evaluating semantic similarity rather than mere lexical overlap, especially for nuanced tasks where phrasing can vary significantly while maintaining meaning. The architectural choice also proved decisive: the encoder-decoder T5 model outperformed the decoder-only GPT-2, likely due to T5's ability to better process and condition on the dual inputs of hate speech and intent. This led us to focus subsequent experiments on T5.

Techniques like prompt augmentation provided modest improvements by enhancing data diversity and model robustness. More significantly, the Intent Contransitive Self-Correction Method demonstrated clear benefits, suggesting that training with both positive and negative examples helps the model develop a more discriminative understanding of intent boundaries.

Conversely, a two-stage fine-tuning approach (general then intent-specific) led to overfitting, indicating that direct, intent-focused training may be more effective for this task. Furthermore, LoRA adapters, despite their parameter efficiency, underperformed compared to full fine-

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**

tuning, suggesting that the full capacity of the model is beneficial for capturing the complexities of this specific generation task.

Conclusions

This research successfully developed and evaluated models for intent-specific counterspeech generation task. We demonstrated that fine-tuned Transformer models, particularly T5, can effectively generate counterspeech aligned with specified intents. Key findings indicate that the T5 architecture is better suited for this conditional task than GPT-2. Furthermore, techniques like prompt augmentation and especially the Intent Contransitive Self-Correction Method enhance performance by improving the model's ability to not differentiate between differnt variations of the input speech and adhere to target intents. While parameter-efficient methods like LoRA and certain staged fine-tuning approaches showed limitations in this context, the overall results underscore the feasibility of generating high-quality, nuanced counterspeech, offering a valuable step towards more effective automated tools for combating online hate.

4.1 Limitations

- **Evaluation Methodology**: Automated metrics (BLEU, ROUGE, BERTScore), while useful, do not fully capture human perception of quality and impact, necessitating comprehensive human evaluation protocols.
- **Informative Counterspeech Generation**: The factual accuracy of current models is constrained by the training dataset; there's a need for methods to integrate verifiable, up-to-date external knowledge.

4.2 Future Directions

- **Knowledge-Augmented Generation**: Explore Retrieval Augmented Generation (RAG) or similar methods to significantly enhance the factual grounding and timeliness of informative counterspeech.
- **Human-in-the-Loop Evaluation**: Implement rigorous human assessment to validate the real-world effectiveness, safety, and nuanced impact of generated counterspeech beyond automated scores.

References

[1] Benesch, S., et al. (2016). Dangerous Speech: A Practical Guide. Dangerous Speech Project.

[2] Chung, M., et al. (2021). Towards a Taxonomy of Counterspeech for Combating Online Hate Speech. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).

[3] Aswini, A. *Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF*. arXiv preprint arXiv:2403.10088, 2024. Available at: https://arxiv.org/html/2403.10088v1#S6

[4] Aswini, A. *IntentCONANv2 Dataset*. Available at: https://huggingface.co/datasets/Aswini123/IntentCONANv2

[5] Teterwak, P., Sun, X., Plummer, B. A., Saenko, K., & Lim, S.-N. (2024). *CLAMP: Contrastive Language Model Prompt-tuning*. arXiv preprint arXiv:2312.01629. Available at: https://arxiv.org/abs/2312.01629