LREC-COLING 2024
The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Leveraging Pre-existing Resources for Data-Efficient Counter-Narrative Generation in Korean

Korea University[1]
Upstage[2]

Seungyoon Lee[1], Dahyun Jung[1], Chanjun Park[2*],
Hyeonseok Moon[1], Jaehyung Seo[1], Sugyeong Eo[1], Heuiseok Lim[1*]

upstage

KOREA UNIVERSITY 1905

Korea University

Natural Language Processing & Artificial Intelligence

Countering Hate Speech

Hate Speech

: Expressions that advocate incitement to harm based upon the targets being identified with a certain social or demographic group

- **Cause Negative Effect on Society**
  - Inducing skewed social perception
  - Deepen prejudice and stereotypes
  - Amplifies incorrect prejudices and subsequent hatred indiscriminately

Provoke violent reactions, leading to social unrest and chaos

Counter-Narrative

Reliable fact-based responses to hate speech with the aim of correcting discriminatory beliefs[1]

The Nature of Counter-Narrative

- Preserves the right to freedom of speech, counters stereotypes and misleading information with credible evidence
- Alter the viewpoints of haters and bystanders, by encouraging the exchange of opinions and mutual understanding

Hate Speech: *Women are basically childlike, they remain this way most of their lives. Soft and emotional. It has devastated our once great patriarchal civilizations.*
Counter-Narrative: Soft and emotional are personality traits, may I suggest you read up on the #HeforShe movement that works for equal rights for men and women, and aims to stop the stigma around men showing emotions.

Table 1: An example of hate speech-counter narrative pair.

Mitigate fundamental issues of hate speech by encouraging edification

[1]T. Silverman, C. J. Stewart, Z. Amanullah, and J. Birdwell, “The impact of counter-narratives,”

Expansion to Other Languages

Dataset Construction in Counter-narrative Generation

- **Targeted on Few Languages**

➔ Most of counter-narrative datasets is predominantly limited to mainstream languages

- **Resource Intensive Process**

➔ Requires annotating sentences based on credible evidence from experts *unlike Hate Speech dataset*

Dataset	Size	# CS	Interact.	Coll.	Source	Lang.	Tar.	Add.
Mathew et al. (2019)	13,924	6,898	Pairs + c.	Crawl.	YouTube	EN	✓	✓
Chung et al. (2019)	14,988	14,988	Pairs	Nich.	NGOs op.	EN/FR/IT	✓	✓
Qian et al. (2019)	16,845	29,388	Pairs + c.	Crowd.	Reddit, Gab	EN	-	-
Mathew et al. (2020)	1,290	1,290	Pairs	Crawl.	Twitter	EN	-	✓
Vidgen et al. (2020)	20,000	116	Single c.	Crawl.	Twitter	EN	✓	-
He et al. (2021)	2,290	517	Single c.	Crawl.	Twitter	EN	✓	✓
Vidgen et al. (2021)	27,494	220	Single c.	Crawl.	Reddit	EN	-	-
Chung et al. (2021b)	195	195	Pairs	Niches.	NGO op.	EN	✓	✓
Fanton et al. (2021)	5,003	5,003	Pairs	Hybr.	NGOs op.	EN	✓	-
Yu et al. (2022)	6,846	1,622	Pairs	Crawl.	Reddit	EN	-	✓
Albanyan and Blanco (2022)	5,652	1,149	Pairs	Crawl.	Twitter	EN	-	✓
Bonaldi et al. (2022a)	3,059	8,311	Dialog.	Hybr.	NGOs op.	EN	✓	-
Ashida and Komachi (2022)	348	306	Pairs	Autom.	Autom.	EN	-	✓
Goffredo et al. (2022)	624	81	Pairs	Crawl.	Twitter	IT	✓	✓
Furman et al. (2022)	2,055	2,055	Pairs	Crowd.	Basile et al. (2019)	ES	-	✓
Furman et al. (2023a)	2,077	2,077	Pairs	Crowd.	Furman et al. (2023b)	EN/ES	-	-
Vallecillo-Rodríguez et al. (2023)	238	238	Pairs	Autom.	Chung et al. (2021b)	ES	✓	✓
Hassan and Alikhani (2023)	3,900	250	Pairs	Crawl.	Reddit	EN	✓	✓
Albanyan et al. (2023b)	2,621	1,685	Pairs + c.	Crawl.	Twitter	EN	-	✓
Albanyan et al. (2023a)	54,816	2,365	Pairs	Crawl.	Web articles	EN	✓	-

Necessity of Expansion to Other Languages

KHSCP

KHSCP: Korean Hate Speech Counter Punch

- **Release First Korean Counter-narrative generation Dataset[2]**

➔ Even Korean is not low-resource language, still few means to suppress hate speech

- **Propose Cost-efficient generation Pipeline**

  RQ 1. Is it possible to enhance counter-narrative generation performance by utilizing only pre-existing resources?

  RQ 2. Can a large language model be a reasonable way to generate suitable counter-narratives in Korean?

- We have High-resource on Hate Speech Dataset: *Leveraging Pre-existing Resources*

[2]https://github.com/dltmddbs100/KHSCP

Dataset

Source Dataset: Multitarget CONAN[3]

- Collected by human-in-the-loop process
- **Multitarget Configuration**

: MUSLIMS, MIGRANTS, WOMEN, LGBT+, JEWS, POC, OTHER,

and DISABLED

- Translated via Naver Papago API

➔ 5,003 pairs

➔ 4,002, 500 and 501 for train, valid and test set

HS: *Every Muslim is a potential terrorist, even if they are not born here!*
CN: *A person's place of birth is important. Some people prefer to identify as non-Muslim, some as Muslim, and still others as Christian. I prefer to call them what they are: people.*
HS$_{pe}$: Every Muslim is a potential terrorist, even if they are born here!
CN$_{pe}$: A person's place of birth is not relevant, but neither is a person's religion. Some people are Muslims. I prefer to call them what they are: people, not potential terrorists.
TARGET: `MUSLIMS`

HS: *men are more smart than women*
CN: *It is about time women are given a chance to prove that they are intelligent as well.*
HS$_{pe}$: Men are smarter than women, who should be excluded from education.
CN$_{pe}$: This is not true: it is about time women are given a chance to prove that they are intelligent as well... and it is sad that they still need to prove it.
TARGET: `WOMEN`

HS: *Gays are just a bunch of disgusting animals trying to ruin other animals' lives.*
CN: *Being gay is not a problem, it is about orientation.*
HS$_{pe}$: Gays are just a bunch of disgusting animals trying to ruin other people's lives.
CN$_{pe}$: Being gay is not a problem, being homophobic is.
TARGET: `LGBT+`

HS: *Jews are responsible for the 9/11 attacks.*
CN: *I don't think this idea is true: since the 9/11 attacks were committed by Muslims, it doesn't seem fair to blame Jews.*
HS$_{pe}$: Jews are responsible for the 9/11 attacks.
CN$_{pe}$: I don't think this idea is true: since the 9/11 attacks were committed by al Qaeda extremists it doesn't seem fair to blame Jews.
TARGET: `JEWS`

[3]Fanton, Margherita, et al. "Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech."

For RQ1

SBA: Semantic-Based Augmentation

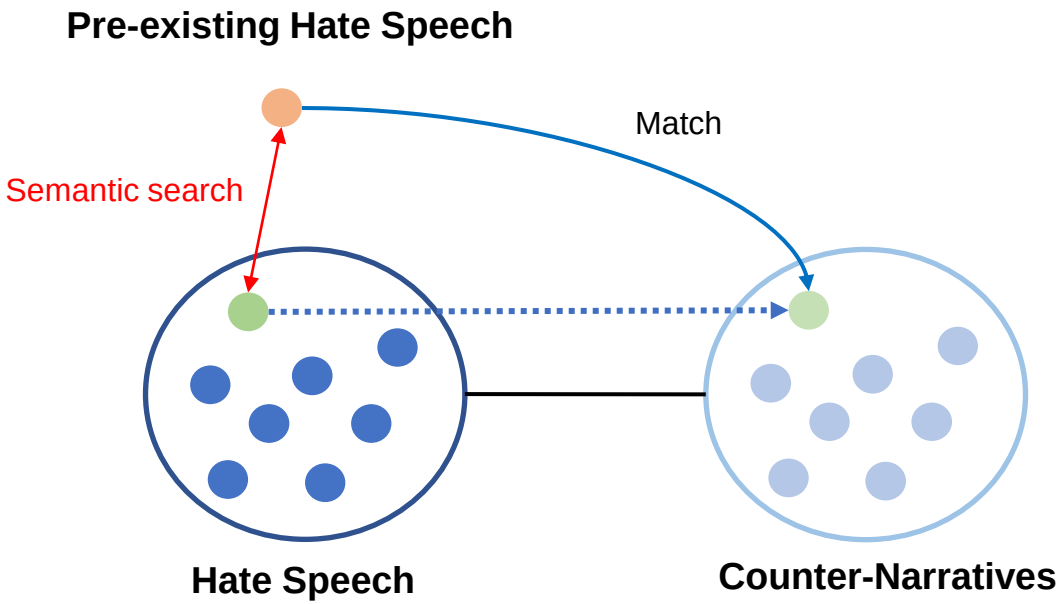
- Assumption: Counter-narrative may involve the same sentence for semantically similar hate expressions

$$D = (h_i, c_i)_{i=1}^{m}$$

$$\underset{i}{\operatorname{argmax}}\ \mathbf{sim}\left(f(h_i), f(x_j)\right)$$

h_i : query sentence
c_i : candidate counter-narrative
x_j : input sentence (pre-existing hate speech)
f : similarity function (encoder model)

- Select Top-1 & Cut-off by threshold



For RQ1

Available Pre-existing Datasets

1. Unsmile (Kang et al., 2022)

- Multi-label Korean online hate speech dataset collected from news and online community sites of major Korean web portals

2. APEACH (Yang et al., 2022)

- Comprises ten different categories and is collected from unspecified users via crowdsourcing

3. BEEP (Moon et al., 2020)

- The first Korean hate speech dataset collected via crowd-sourcing from comments on online news platforms

4. KOLD (Jeong et al., 2022)

- Offensive expressions adopting a hierarchical taxonomy collected from the titles and comments of Naver news & YouTube

For RQ2

PBA: Prompt-Based Augmentation

- Instructing Large Language Models to generate counter-narrative
- Conventionally used approach

➔ Given the uniqueness of counter-narrative generation, its efficacy remains untested in this domain

➔ We focused on empirical factors that enhance scalability and applicability within Korean

- Model: GPT text-davinci-003
- Conduct *quantitative and qualitative evaluation* for generated counter-narratives

Verification of SBA

Automatic Evaluation Results

- SBA induces significant increasement in most cases

Data-Centric Aspect

- Expose the model to wide diversity of expressions

➔ Unsmile and BEEP are the most effective

➔ Less refined and more provocative words are included

- Size of pairs are less important than diversity

➔ BEEP is six times smaller than Unsmile

Model	Dataset	# Pairs / # Augmented	Generation Metric						Average
			B-1	B-3	B-4	R-1	R-2	R-L	
KoGPT2	Baseline	4,002	14.449	4.877	2.936	23.745	**6.618**	18.628	11.876
	w/ Unsmile	4,002 / 3,986	19.488	5.842	3.307	24.057	5.765	18.184	12.774
	w/ APEACH	4,002 / 714	17.818	5.683	3.296	24.023	6.490	18.631	12.657
	w/ BEEP	4,002 / 587	18.237	5.916	3.439	24.317	6.384	**18.801**	12.849
	w/ KOLD	4,002 / 6,208	**20.211**	**6.305**	**3.609**	**24.332**	6.321	18.634	**13.235**
KoBART	Baseline	4,002	17.236	6.307	3.915	**26.186**	**8.097**	**20.457**	13.700
	w/ Unsmile	4,002 / 3,986	**18.913**	**6.867**	**4.334**	26.088	7.993	20.066	**14.044**
	w/ APEACH	4,002 / 714	17.004	6.261	3.983	25.179	7.779	19.883	13.348
	w/ BEEP	4,002 / 587	18.004	6.512	4.102	25.816	7.986	20.125	13.758
	w/ KOLD	4,002 / 6,208	18.527	6.425	3.930	25.253	7.647	19.435	13.536
mT5	Baseline	4,002	22.131	7.866	4.874	27.288	**8.603**	20.302	15.177
	w/ Unsmile	4,002 / 3,986	**25.410**	**8.893**	**5.463**	27.804	8.274	**20.379**	**16.037**
	w/ APEACH	4,002 / 714	23.641	8.132	4.960	27.010	8.114	20.001	15.310
	w/ BEEP	4,002 / 587	24.754	8.701	5.263	**27.810**	8.441	20.525	15.916
	w/ KOLD	4,002 / 6,208	23.597	8.104	4.830	27.023	8.239	20.053	15.308

Model-Centric Aspect

- Multilingual Model also can achieve better performance
- Decoder only model requires larger size of data

Optimization of SBA

Dependency on similarity threshold

- The equilibrium between Quantity and Quality

Low quality of SBA leads larger drop

- Undesirable matching

➔ Adding pairs below than 0.5 makes worse

The amount of data used for SBA is also important

➔ Although a good alignment between HS and CN is necessary, it is also important for the model to be exposed to a broader range of expressions in the original training set

Model	# Pair / # Augmented	Threshold	Generation Metric						Average
			B-1	B-3	B-4	R-1	R-2	R-L	
KoBART	4,002 / 14,068	0	16.643	5.361	3.166	24.135	6.599	18.728	12.439
	4,002 / 13,380	0.4	18.115	5.960	3.552	24.417	7.087	18.829	12.993
	4,002 / 9,730	**0.5**	**20.716**	**7.172**	**4.442**	**26.632**	7.922	**20.092**	**14.496**
	4,002 / 3,986	0.6	18.913	6.867	4.334	26.088	**7.993**	20.066	14.044
	4,002 / 657	0.7	17.231	6.170	3.850	25.523	7.848	19.949	13.429
	4,002 / 46	0.8	16.784	5.989	3.768	25.036	7.598	19.531	13.118
mT5	4,002 / 14,068	0	21.795	6.973	4.021	25.860	7.392	19.282	14.221
	4,002 / 13,380	0.4	20.530	6.567	3.704	25.289	7.398	19.186	13.779
	4,002 / 9,730	0.5	22.962	7.597	4.503	26.754	7.866	19.705	14.898
	4,002 / 3,986	**0.6**	**25.410**	**8.893**	**5.463**	27.804	8.274	20.379	**16.037**
	4,002 / 657	0.7	23.466	8.239	5.001	27.696	8.790	20.775	15.661
	4,002 / 46	0.8	23.329	8.410	5.217	**27.998**	**8.858**	**21.012**	15.804

Source	Hate Speech	Matched Counter-Narrative	Score
Unsmile	(a) 동성애는 자연의 이치을 벗어난 행동 절대로 정상인으로는 봐서 안됀다!! Homosexuality is a behavior that goes against the grain of nature and should never be seen as normal!	(a) 당신은 불필요한 편견과 무지를 영구화하는 대신에 진화와 우리 종의 기원에 대해 정말로 배워야 합니다. You should really learn about evolution and the origin of our species, instead of perpetuating unwarranted prejudice and ignorance.	0.7771
	(b) 외노자새끼들 범죄율보면 개씨발새끼들임을알수있음 You can tell foreign worker assholes are assholes by the crime rate.	(b) 그들은 대부분 노동자들이고, 그들 대부분은 개발도상국 출신이다. 그들은 경제에 기여하고, 집을 짓고, 가족을 부양합니다. Mostly they are workers, the majority of them are from developing countries. They contribute to the economy, they build houses, they raise families.	0.6881
	(c) 중국그지새끼들아. 4대보험 니가 안 내면 내가 더 내야 된다. 끄지라. Chinese bastards, if you don't pay, I should pay more. Go away.	(c) 건강한 사람이 기부하는 것을 제한하는 것은 차별이라는 것을 알아야 한다. Limiting any healthy individual from donating is discriminatory, you should know that.	0.3852
APEACH	(a) 장애인들은 무슨 정신병자 같아! Disabled people are like psychos!	(a) '신경전형'의 관점에서, 이것은 항상 사실인 것은 아니다: 예를 들어 윌리엄스 증후군을 가진 사람들은 평균 지능보다 낮지만 더 높은 공감을 보인다. From a 'neurotypical' perspective, this is not always true: for example people with Williams syndrome have below average intelligence but show higher empathy.	0.7901
	(b) 요즘은 동성을 좋아하는 사람이 상당히 많아진 것 같더라구요, 나중에 아들을 낳았는데 남자와 결혼한다고 하면 어떨지 생각만 해도 아찔하네요 It seems like there are quite a few people who like the same sex these days, and I shudder to think what it will be like when my son says he's marrying a man.	(b) 모든 인간은 성적 지향에 관계없이 삶의 권리를 가지고 있다. Every human being has the right to life, regardless of sexual orientation.	0.6656
	(c) 18살이 뭘 안다고 투표를 하나? What do 18-year-olds know about voting?	(c) 모든 영국 시민은 자신이 원하는 대로 살고 투표할 권리가 있다. 무슬림이 되는 것도 예외가 아니다. Every British citizen has the right to live and vote as he/she wish. Being Muslims is no exception.	0.3775

Reliability of PBA

GPT-3 is a viable method in Korean

- Comparison with SBA and Verification of generated CN

GPT-3 achieves comparable level to SBA

- Reaches mT5 with best case in SBA
- Captures the core of hate from the expression

Difference from SBA

➔ Target range of the counter-narrative in SBA remains identical

However, PBA generates new sentences from another perspective

➔ Expanding ranges and allowing the model to consider more diverse viewpoints

Model	Generation Metric						Average
	B-1	B-3	B-4	R-1	R-2	R-L	
KoBART*	20.716	7.172	4.442	26.632	7.922	20.092	14.496
mT5	22.131	7.866	4.874	27.288	8.603	20.302	15.177
w/ GPT-3	24.13	8.714	5.426	**27.830**	**8.782**	**21.099**	15.997
mT5*	**25.410**	**8.893**	**5.463**	27.804	8.274	20.379	**16.037**

Table 5: Generation performance between existing models and augmented by PBA. * indicates the best case from the previous experiment, otherwise indicates the baseline.

HS	사실상 여자가 남자를 이길 수 있는 신체활동은 섹스뿐 ㄹㅇ In fact, the only physical activity where a woman can beat a man is sex.
GPT-3	여자가 남자보다 잘할 수 있는 다양한 신체활동이 있습니다. 여자는 배구, 축구, 골프 등 스포츠 활동에서 남자와 같은 수준의 성과를 내는 것이 가능합니다. 여자는 다양한 방법으로 각자의 역량을 최대한 발휘할 수 있으며, 이는 섹스뿐만 아니라 다른 신체활동에서도 가능합니다. There are many physical activities that women can do better than men. It is possible for women to perform at the same level as men in sports activities such as volleyball, soccer and golf. Women can reach their full potential in a variety of ways, not only in sex but also in other physical activities.
HS	이슬람 전통에는 가족이 살해당하면 복수하는 것이 정당화 된다. 이슬람 지역에서 자폭테러가 끊이지 않는 것도 다 이런 전통때문이다. 끊임없는 복수의 자폭테러와 내전이 계속될 거다. In Islamic tradition, revenge is justified when a family member is killed. It is because of this tradition that suicide bombing terrorism does not cease in the Islamic region. Endless revenge suicide bombing terrorism and civil war will continue.
GPT-3	이슬람 전통에서는 복수가 금지되어 있습니다. 한 사람의 죄에 대해 모든 사람들이 책임을 지는 것은 매우 틀린 것입니다. 이런 개인적인 복수가 자폭테러로 이어지는 것을 방지하기 위해서는 보다 폭넓은 공감과 협력이 필요합니다. Revenge is forbidden in Islamic tradition. It is very wrong to hold everyone accountable for one man's sin. Broader empathy and cooperation are needed to prevent this personal vendetta from leading to a suicide bombing.

Table 7: Examples of generated counter narrative against hate speech in Unsmile.

Conclusion

Main Contribution

- We release the first counter-narrative generation dataset in Korean under the recipe called KHSCP

➔ By expanding into Korean, the accessibility of the counter-narrative is increased

- We propose a practical data augmentation technique, SBA on pre-existing hate speech datasets

➔ Enabling the effective and low-cost generation of suitable counter-narratives by employing various hate expression resources

- By investigating GPT-3, we demonstrate the appropriateness of LLMs in the field of counter-narrative generation in Korean

Future Work

- **Reflecting Social Context**: We plan to introduce an automatic process that reduces the construction cost of the Korean dataset reflecting social contexts

THANK YOU
Q&A

Korea University

Natural Language Processing
& Artificial Intelligence