Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech

Yi-Ling Chung[1,2], **Serra Sinem Tekiroğlu**[1], **and Marco Guerini**[1]

[1]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
[2]University of Trento, Italy

ychung@fbk.eu, tekiroglu@fbk.eu, guerini@fbk.eu

Abstract

Tackling online hatred using informed textual responses – called counter narratives – has been brought under the spotlight recently. Accordingly, a research line has emerged to automatically generate counter narratives in order to facilitate the direct intervention in the hate discussion and to prevent hate content from further spreading. Still, current neural approaches tend to produce generic/repetitive responses and lack grounded and up-to-date evidence such as facts, statistics, or examples. Moreover, these models can create plausible but not necessarily true arguments. In this paper we present the first complete knowledge-bound counter narrative generation pipeline, grounded in an external knowledge repository that can provide more informative content to fight online hatred. Together with our approach, we present a series of experiments that show its feasibility to produce suitable and informative counter narratives in in-domain and cross-domain settings.

1 Introduction

Standard approaches for online hate mitigation generally rely on content moderation, ranging from deletion of hate content and suspension of user accounts to shadow banning. However, these approaches may draw limits on free speech and diverse opinions. An alternative approach is to directly intervene in the conversation with counter narratives. A Counter Narrative (CN) is a non-negative response to a Hate Speech (HS), targeting and contradicting extreme statements with fact-bound arguments or alternative viewpoints (Benesch, 2014; Schieb and Preuss, 2016). Such strategy seeks to de-escalate the conversation, disengage from hateful sentiment and encourage mutual understanding through exchange of opinions. Many Non-Governmental Organizations (NGOs) specialized in hate countering are already adopting this approach by training operators to compose counter narratives. According to NGO guidelines[1], proper CNs should also include credible evidence.

HS:	*The world would be a better place without Muslims. They are only killing and raping our children.*
CN1:	*Evidence for this? This is not true. How can you say this about an entire faith?*
CN2:	*On the contrary, most children abuse is operated by people they know: a relative, family friends, sports coach, someone in a position of trust and authority. Besides, Muslims help people - A Muslim woman rushed to help the victims of a triple stabbing in Manchester on New Year's Eve.*
CN3:	*You are truly one stupid backwards thinking idiot to comment on Muslims like that.*

Table 1: An example hate speech with appropriate counter narratives (CN1 and CN2) and an inappropriate one (CN3) containing abusive language.

In Table 1, we present an HS along with several CNs. Although CN1 and CN2 are both appropriate responses, not all appropriate CNs are equally effective (Silverman et al., 2016; Tuck and Silverman, 2016). CN2 is expected to be more effective as it is tailored to the HS and demonstrates contextualized and persuasive statements with supporting facts, knowledge, and logical reasoning (Habernal and Gurevych, 2016) rather than a simple generic request for evidence as in CN1. Conversely, CN3 is considered as a natural but inappropriate and aggressive response.

In this context, scaling manual response to massive amounts of online hatred is an undeniably challenging task for NGO operators. Recently, studies have started to focus on data and strategies for the automatic generation of counter narratives in an end-to-end fashion (Chung et al., 2019; Qian et al., 2019; Tekiroğlu et al., 2020; Chung et al., 2020). However, the seminal studies, based on training

[1] For example: https://getthetrollsout.org/stoppinghate

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 899–914
August 1–6, 2021. ©2021 Association for Computational Linguistics

models using only the HS-CN data samples, do not directly address the required properties for proper and effective CNs such as providing credible evidence and knowledge with facts, statistics or examples. While such models possess knowledge to a certain extent through utilizing existing generative language models (LMs) such as GPT-2 (Radford et al., 2019), this knowledge (e.g., events or news) will get out-dated over time. Furthermore, such models are able to create plausible but not necessarily true arguments - a problem known as content hallucination - (Zellers et al., 2019; Solaiman et al., 2019). A hallucinated example of a fact-bound CN, generated by GPT-2 model fine-tuned as done by Tekiroğlu et al. (2020), is provided in Table 2. Therefore, in this paper we investigate the generation of knowledge-bound counter narratives, which had never been studied yet to the best of our knowledge. We hypothesized that knowledge infusion can not only leverage generating informative CNs but also handling hate speech from unseen domains (i.e. hate targets) without further training (e.g., train on Islamophobia and test on Antisemitism).

HS: *Muslims are invading our country*
CN: *Actually, there were 16,938,000 Muslims in the UK in 2016, so if you exclude London, that is actually increasing by 2%, which doesn't seem very significant.*

Table 2: Hallucinated CN generated by GPT-2 that is fluent and credible (according to Office for National Statistics, the Muslim population is just above 3M).

To this end, we explore methodologies to generate informative CNs using external knowledge. In particular, we propose an extension of knowledge-grounded generation approaches by adopting an intermediate step where we generate keyphrases for retrieving needed knowledge. So, we first train a counter narrative keyphrase generator, then the generated keyphrases are employed for selecting relevant knowledge sentences. Finally, pre-trained LMs are fine-tuned on the relevant knowledge sentences, together with the HS input, to produce knowledge-augmented CNs. Our extensive experiments on CN generation, including both automatic and expert evaluation, demonstrate that the presented approach produces more specific and tailored responses both for in-domain and zero-shot cross-domain configurations as compared to other approaches, such as standard LMs, that are simply fine-tuned for the task without the use of external knowledge.

As our main contribution, we show that: (i) external knowledge can boost informative CN generation, (ii) keyphrase generation improves the quality of retrieved documents, (iii) silver knowledge is utilizable for the task when no gold knowledge is available, (iv) knowledge-bound models are advantageous while tackling zero-shot cross-domain generation especially if (v) using injection of knowledge in large pre-trained LMs.

2 Related Work

In this section we review three main research topics that are relevant for fighting hatred online: (i) studies on CN effectiveness in hate countering, (ii) counter-argument generation and (iii) knowledge-guided generation.

Hate countering. Employing counter narratives has shown to be an effective strategy in hatred intervention on social media platforms. Studies have focused on identifying successful counter narratives (Benesch et al., 2016a,b), evaluating their efficacy (Schieb and Preuss, 2016; Silverman et al., 2016; Ernst et al., 2017; Munger, 2017), and analyzing counter speaker accounts characteristics (Mathew et al., 2018). In particular, by analyzing conversations from Twitter, Wright et al. (2017) show that some arguments among strangers induce favorable changes in discourse and attitudes.

Counter-argument generation shares similar objectives as CN generation, i.e., to produce the opposite or alternate stance of a statement, but the latter faces peculiar difficulties such as the absence in HS of explicit or well-structured 'arguments' (e.g., "*Islam is a disease*") and the limited amount of data available for training. Studies usually focus on domains with large discussions, e.g., politics (Hua and Wang, 2018) and economy (Le et al., 2018). The closest work to ours is counter-argument generation with external knowledge augmentation by Hua et al. (2019). Our approach differs from theirs in three aspects: (i) we explore generating queries to extract knowledge for grounding CN with, (ii) pre-trained generative models are utilized for leveraging the knowledge present, (iii) our approach requires less manipulation over knowledge.

Knowledge-guided generation. There is a growing interest in exploiting external knowledge to generate informative responses for applications such as dialog systems (He et al., 2017; Young et al., 2018) and question answering (Das et al., 2017; Saha et al., 2019). Previous approaches inject knowledge through topic phrases (Fan et al., 2019), structured knowledge graphs (Zhou et al., 2018) and unstructured texts (Dinan et al., 2019; Hua et al., 2019).

3 HS-CN Dataset

To the best of our knowledge, there is no high-quality hate speech - counter narrative dataset available yet where CNs are explicitly paired with relevant knowledge. Since constructing such dataset with a decent-size would be too costly and out of the scope of the present paper[2], we resort to a "reverse-engineering" strategy such that we automatically paired relevant knowledge with an already existing high quality CN dataset. We chose CONAN (Chung et al., 2019), which is a dataset niche-sourced to expert NGO operators offering high quality CNs, and the best and most diverse material among the other CN datasets (Tekiroğlu et al., 2020). CONAN consists of 6645 English pairs of HS-CN including: 1288 original pairs, 2576 pairs where two paraphrases of the original HS are paired with the original CN, and 2781 translated pairs from French and Italian. The English data is split into 4069/1288/1288 samples for train/dev/test.

4 Architecture

Our architecture, illustrated in Figure 1, consists of a knowledge retrieval module that retrieves sentence-level relevant knowledge, and a generation module that generates a counter narrative. Specifically, the knowledge retrieval module first prepares variants of a query Q for a given hate speech HS using two strategies: query extraction (Q_{hs}) and automatic query generation (Q_{gen}). Then, the obtained queries are employed to search for relevant knowledge articles via a search engine. Finally, it uses a sentence selector to filter and rank the most relevant sentences as the relevant knowledge (KN) from the retrieved articles. For the counter narrative generation module, we fine-tuned several LMs that take a HS and the ranked knowledge sentences KN as input and output a corresponding counter narrative.

5 Knowledge Retrieval Module

The knowledge retrieval module in the architecture incorporates a knowledge repository, query construction sub-module, and a knowledge sentence selection sub-module.

5.1 Knowledge Repository

Previous approaches on introducing external knowledge for dialog generation have exploited unstructured and structured knowledge. Since no structured knowledge is available for the hate speech domain, we rely on unstructured textual knowledge in the format of articles, which allows for updating the knowledge repository easily. Considering that the proliferation of HS is also triggered with target-related events (e.g., terrorist attacks), being able to update the knowledge, such as news articles, would let us produce proper CNs that contain the latest statistics or evidence from the current events.

We include Newsroom (Grusky et al., 2018) and WikiText-103 (Merity et al., 2017) to our knowledge repository. WikiText-103 is a large collection of 28,595 full Wikipedia articles covering over 103 million words. Newsroom consists of 1.3 million articles extracted from major news publications between 1998 and 2017, featuring over 6.9 million words.

5.2 Query Construction

To construct comprehensive and proper queries to search for relevant knowledge for the data pairs, we applied two strategies: (i) query extraction and (ii) query generation. In both strategies, the query is composed of keyphrases that can be defined as the important and topical phrases from a text (Turney, 2000).

Query extraction. We extracted keyphrases from CONAN dataset using Keyphrase Digger (Moretti et al., 2015), a multilingual keyphrase extraction system that uses statistical measures and linguistic information, and is proven to be one of the best systems for unsupervised settings[3]. Following the knowledge retrieval strategy using input argument by Hua et al. (2019) for counter argument generation, we first obtained the HS keyphrases to construct the initial query Q_{hs}. However, HSs from CONAN mostly contain hateful and simplistic

[2] Obtaining access to a pool of trained NGO operators is very complicated, furthermore keeping track of their search activity and the material they used during CN production would require long and complex data collection sessions that might span several months.

[3] Keyphrase Digger is a new implementation of KX (Pianta and Tonelli, 2010) with several improvements, and it was ranked the best performing unsupervised system on task 5 of SemEval 2010 evaluation campaign (Moretti et al., 2015).

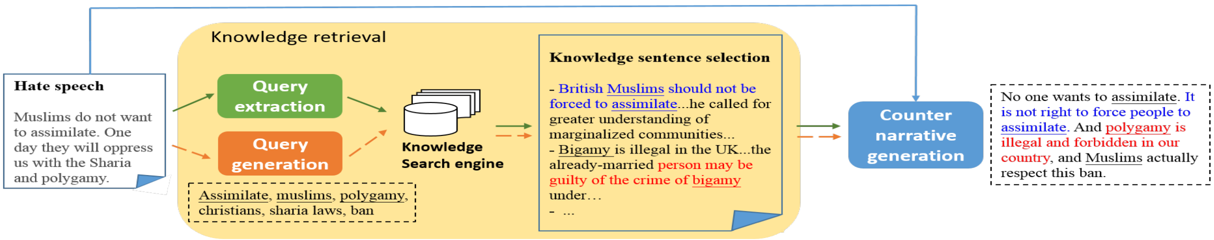
Hate speech
Muslims do not want to assimilate. One day they will oppress us with the Sharia and polygamy.
Knowledge retrieval
Query extraction
Query generation
Knowledge Search engine
Assimilate, muslims, polygamy, christians, sharia laws, ban
Knowledge sentence selection
- British Muslims should not be forced to assimilate...he called for greater understanding of marginalized communities...
- Bigamy is illegal in the UK...the already-married person may be guilty of the crime of bigamy under…
- …
Counter narrative generation
No one wants to assimilate. It is not right to force people to assimilate. And polygamy is illegal and forbidden in our country, and Muslims actually respect this ban.

Figure 1: Architecture of knowledge grounded generation with extracted (green solid arrow) and generative (dotted arrow) queries (topical phrases) that are exploited to retrieve relevant knowledge. The knowledge sentences extracted together with input HS are fed to CN generation. We give the example of generative approach.

phrases in comparison to the input arguments used by Hua et al. (2019) that can be rich in content[4]. Therefore, in the HS-CN scenario, we hypothesize that the keyphrases from Q_{hs} alone would not be sufficient for relevant knowledge search especially for mapping the knowledge onto training data.

To this end, we also extracted keyphrases from CN together with HS to increase the possibility that the retrieved knowledge sentences contain pieces of information found in the ground truth. Hence, the second query $Q_{hs \cup cn}$ contains CN keyphrases for the relevancy to the target CN and HS keyphrases for preserving the hate context. We investigated the effects of various keyphrase query configurations in terms of HS relevancy and $Q_{hs \cup cn}$ is proven to be the best configuration (See Appendix A.1 for more details.).

Query generation. Since the best query configuration $Q_{hs \cup cn}$ cannot be available at test time, we need a way to obtain keyphrases that serve as CN cues for searching knowledge sentences during the CN generation. To this end, we built a query generation model that takes HS as input and outputs a comma-separated list of CN keyphrases, which is then used as Q_{gen}. Our aim is to obtain an approximation of $Q_{hs \cup cn}$ via $Q_{hs \cup gen}$ at the test time.

The model is trained using Transformer (Vaswani et al., 2017) architecture as it has obtained state-of-the-art performances for generation tasks (Dinan et al., 2019; Ghazvininejad et al., 2018). For the training data, we used CONAN dataset and discarded the CNs that are less than 10 words, since they are usually generic, poor in terms of argumentative content and cannot provide a meaningful search (e.g., "*No they are not - prove this?*", "*What does that even mean?*", "*Any evidence?*"). Accordingly, we kept 4038/1257/1257 instances for train/dev/test set. The train set includes the pairs marked as original in the dataset, and all translated pairs from French and Italian; the dev set consists of one paraphrase of each original HS and its CNs; and the test set contains the rest of the paraphrased HSs. The training inputs are represented as HS $[HS\_end\_token]$ KP $[KP\_end\_token]$, where KP is the list of keyphrases extracted from the gold CN.

The model has been trained following the configuration of the base model in (Vaswani et al., 2017): with 6 transformer layers, 8 transformer heads, embedding size of 512, hidden size of 2048, dropout rate of 0.1, batch size of 64 for 100 epochs. The training time lasted around 7 hours. All experiments in this paper have been conducted on a Nvidia Tesla V100 GPU. For decoding, we used nucleus sampling (Holtzman et al., 2020) with a p value of 0.9.

We report keyphrase generation results in terms of BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) against the keyphrases extracted from gold test CNs. We obtained a score of 0.162 for BLEU-2 and a score of 0.353 for ROUGE-L. Although both scores can be considered as low, this is due to the open ended nature of the set of possible CN keyphrases for a given HS. Example queries for a single pair, extracted from its HS, its CN, and generated with the keyphrase generation model are shown in Table 3.

5.3 Knowledge Sentence Selection

We use Solr[5] to index the articles and retrieve those relevant to a given query based on the similarity between the articles and the query using BM25 (Robertson et al., 1995). Once the queries have been obtained either through extraction or generation, they are presented to Solr for retrieving

[4] e.g., an argument "*A universal basic income will help the labor market adapt to inevitable disruptions caused by advancements in automation and artificial intelligence.*" vs. HS "*Islam is a disease*".

[5] https://lucene.apache.org/solr/

HS	CN	Query	Knowledge sentences (KN)
Islam is a disease.	Like Christianity or any other, islam is a religion of tolerance. Disease does not discriminate on religious grounds.	Q_{hs}: islam, disease	(i) Do Muslims want to heal from the disease...? (ii) Being infected by religious extremism is like being infected by a disease...
		Q_{gen}: islamic law, god, christians	(i) Islamic law is to create an environment...submission to God. (ii) Certain areas of the Muslim world have always been home to large populations of Christians...
		Q_{cn}: tolerance, christianity, discriminating	(i) Islam is a 1400-year-old religion that preaches tolerance...like Christianity and... (ii) Disease does not discriminate...on religious grounds...

Table 3: Examples of KN retrieved using queries extracted from HS (Q_{hs}), generated (Q_{gen}) and created from both HS and CN keyphrases (Q_{cn}).

the 25 top-ranked articles. Next, we used spaCy sentence segmentation[6] to split an article into sentences. Similar to Zhang et al. (2020), given a query Q we score each sentence x_i in the set of articles D independently, using ROUGEL-F1 (Lin, 2004) as in Equation 1.

$$s_i = rouge(x_i, Q), \forall i \in D \qquad (1)$$

In the final step, we distilled the knowledge by keeping the top 5 knowledge sentences that have the highest scores among 25 top-ranked articles. Instead of a more stringent filtering, such setting has been applied to grant a better variety of source articles and corresponding distilled sentences. We refer to such automatically associated sentences as "silver knowledge".

6 Counter Narrative Generation Module

Large pretrained LMs require less amount of high-quality data to be fine-tuned on downstream tasks while providing strong performances and they already store large amount of factual and commonsense knowledge from their training data (Petroni et al., 2019). To this respect, we built the following models: (1) **GPT-2$_{KN}$**, obtained by fine-tuning GPT-2 on CONAN data paired with KN; (2) **GPT-2$_{KN,MT}$**, by fine-tuning GPT-2$_{KN}$ in a multi-task learning fashion for learning to distinguish CNs from HS as next utterances; (3) **XNLG** (Chi et al., 2020) for its ability to copy information to the output (in our case the retrieved KN to be copied to the CN). We expect all three models to attend over the HS and retrieve KN and look for the relevant snippets to be recovered while generating a CN.

6.1 Models

The training HS-CN pairs are represented as HS $[HS\_end\_token]$ KN $[KN\_end\_token]$ CN $[CN\_end\_token]$. Each model is trained with $Q_{hs \cup cn}$ and then tested on Q_{hs}, Q_{gen}, and $Q_{hs \cup gen}$. We also tested the models with $Q_{hs \cup cn}$ to define an oracle scenario with an upperbound performance when the data can only be paired with silver knowledge.

GPT-2$_{KN}$. We fine-tuned the GPT-2[7] medium model for 3 epochs with a batch size of 2048 tokens. We used Adam optimizer with a learning rate of 5e-5. At inference time, responses were generated employing nucleus sampling with a p value of 0.9, conditioned on HSs and corresponding KN.

GPT-2$_{KN,MT}$. Since we noticed that GPT-2 occasionally produces responses that contain fragments of abusive language, we combined the language modeling objective with a next-sentence prediction objective for fine-tuning GPT-2 in a multi-task setting, inspired by Wolf et al. (2018). Next-sentence prediction adopts a linear classification layer added to the last layer of the transformer language model and then applies a cross-entropy loss to classify a proper next response to the input HS from 2 distractors randomly selected from HS. We used Adam optimizer with a learning rate of 5e-5 and empirically fine-tuned it for 1 epoch and the same sampling strategy as GPT-2$_{KN}$ has been applied.

XNLG is a pre-trained Transformer-based language model trained on Wikipedia dumps with two relevant objectives for our task: to obtain contextual representations and to recover a given input. We fine-tuned XNLG[8] for generating counter narratives on all layers with a batch size of 10 for 100 epochs. We used Adam optimizer with a learning rate of 1e-4. We tokenized and removed accent from the entire dataset and applied the same BPE

[6] https://spacy.io/universe/project/spacy-sentence-segmenter

[7] https://github.com/huggingface/transformers

[8] https://github.com/CZWin32768/XNLG

codes used by Chi et al. (2020). For KN and CN we kept the first 256 tokens, while setting the HS to 70 tokens, which is the maximum length of hate speech in the dataset. We experimented with various decoding methods and adopted beam search with a beam-width of 3 for the best performing setting (details in Appendix A.2).

Baselines used for comparison are: (1) non-pretrained **Transformer** without knowledge using the same hyper-parameters as keyphrase generation model; (2) **GPT-2** without knowledge following the same configuration as GPT-2_{KN}; (3) **Candela** (Hua et al., 2019), an LSTM-based state-of-the-art knowledge-driven architecture for argument generation. Since CONAN is relatively small, we hypothesize that a pre-training procedure[9] on data from a similar task (argument generation) can be beneficial for generalization and porting knowledge. Thus, we first pre-trained Candela architecture on argument generation dataset (Hua et al., 2019), following the configuration described in the paper. We then fine-tuned the model for 20 epochs on CONAN with KN using Q_{hs} as it is done in the original setting of Candela.

6.2 Results for the Silver Knowledge Test Set

We report BLEU-2 (B-2) and ROUGE-L (R-L) scores for all proposed models and baselines in Table 4 on the test split of CONAN that we automatically paired with silver knowledge using various queries. We also measure the capability of each model to produce *novel* responses with respect to the training data by Jaccard similarity (Wang, 2018), and *diverse* responses for the given input by repetition rate (RR) (Cettolo et al., 2014).

Among our models, GPT-2_{KN} yields the highest B-2 and XNLG the highest novelty, diversity, and R-L. The notably improved novelty achieved by knowledge-grounded models indicates the benefits of adding knowledge on producing CNs, in comparison to the baselines - particularly Transformer TRF. On the other hand, the quantitative performance of XNLG does not reflect its true performance in terms of quality. A quick glance at the output CNs showed that XNLG model copies almost everything from KN to the output instead of a proper CN generation, increasing the novelty and diversity scores. The issue can easily be observed from the average numbers of words and sentences in the XNLG output in comparison to the outputs of the other models presented in Table 4. GPT-$2_{KN,MT}$ falls behind among our models in terms of RR, B-2, and R-L, still providing a competitive novelty. Regarding Candela, while it obtained similar performances to our models in terms of R-L and B-2, the generation is repetitive and less novel.

As for the testing with different query types, $Q_{hs \cup gen}$ induces more novel responses than $Q_{hs \cup cn}$ and Q_{hs}. While XNLG yields the highest novelty with Q_{hs} (0.824), it can be explained again with the problem of copying the whole KN, which is more varied due to the less restrictive search using only HS.

The oracle query $Q_{hs \cup cn}$, in which we deliberately provide the best knowledge possible through the keyphrases containing also from the gold CN, yields the best R-L scores among the query variations of knowledge-grounded models. Among all the models, $Q_{hs \cup cn}$ also leads to the best B-2 through GPT-2_{KN}, and the best R-L through XNLG, as we have anticipated. Finally, $Q_{hs \cup gen}$ outperforms Q_{hs} and Q_{gen} over most metrics, hinting at the advantages of using generated queries together with hate context for silver-knowledge retrieval.

We have also conducted complementary experiments by taking into consideration the design choices and the various phenomena in our study. Since in our test set, in line with CONAN, a HS can be paired with more than one CN, Q_{hs} would retrieve the same KN for all the target CNs of the same input HS. Contrarily, we obtain a different set of KN using queries Q_{gen}, $Q_{hs \cup gen}$ and $Q_{hs \cup cn}$ for each target CN. Therefore, we also report an evaluation on unique HS-CN pairs, where a single target CN has been randomly chosen for each HS, among all query types in Appendix A.4. Finally, to simulate Candela configuration (that uses only Q_{hs}) also with the other models, we run an additional set of experiments where we used Q_{hs} for retrieving the knowledge for training samples. The results are reported in Appendix A.3.

6.3 Results for the Gold Knowledge Test Set

To isolate the effect of the knowledge retrieval strategies from the knowledge-grounded generation performances, we conducted a second evaluation on a newly crafted test set paired with gold standard knowledge. In this evaluation, in addition to stereotypical islamophobic in-domain (i.e.

[9] We also trained Candela from scratch on CONAN but decided not to proceed with this setting for the poor performance.

							KN overlap (ngram)		
Models	Nov.	RR	B-2	R-L	#Word	#Sent.	1	2	3
without knowledge									
TRF	0.467	7.72	0.082	0.094	21.47	1.70	-	-	-
GPT-2	0.688	9.04	0.045	0.100	15.95	1.35	-	-	-
Train$_{cn}$	-	3.91	-	-	21.79	1.87	0.307	0.054	0.016
with knowledge									
Candela (Q_{hs})	0.692	21.87	0.040	0.098	23.85	2.47	0.173	0.008	0.001
GPT-2$_{KN}$									
w/ Q_{hs}	0.723	8.13	0.082	0.094	15.60	1.32	0.258	0.023	0.008
w/ Q_{gen}	0.728	7.48	0.067	0.091	12.75	1.17	0.260	0.050	0.019
w/ $Q_{hs \cup gen}$	0.735	6.30	0.085	0.103	15.35	1.59	0.358	0.068	0.024
w/ $Q_{hs \cup cn}$	0.727	7.17	**0.166**	0.110	13.10	1.16	0.282	0.058	0.022
GPT-2$_{KN,MT}$									
w/ Q_{hs}	0.744	11.69	0.050	0.090	13.35	1.17	0.269	0.049	0.017
w/ Q_{gen}	0.731	10.37	0.052	0.092	13.34	1.14	0.253	0.044	0.017
w/ $Q_{hs \cup gen}$	0.747	7.59	0.091	0.090	16.91	1.26	0.269	0.033	0.009
w/ $Q_{hs \cup cn}$	0.731	9.56	0.048	0.107	13.05	1.13	0.276	0.057	0.023
XNLG									
w/ Q_{hs}	**0.824**	14.42	0.073	0.084	55.51	3.71	0.841	0.650	0.558
w/ Q_{gen}	0.819	6.88	0.097	0.084	55.64	3.64	0.849	0.656	0.558
w/ $Q_{hs \cup gen}$	0.812	6.98	0.074	0.089	57.58	3.00	0.828	0.579	0.475
w/ $Q_{hs \cup cn}$	0.819	**5.69**	0.076	**0.116**	55.69	3.42	0.840	0.631	0.529

Table 4: Results of CN generation with silver knowledge. We report novelty (Nov.), RR, BLEU-2 (B-2), ROUGE-L (R-L), KN overlap with generation and the average amount of words and sentences per generation.

	in-target				cross-target			
	Nov.	RR	B-2	R-L	Nov.	RR	B-2	R-L
TRF	0.30	7.57	0.014	0.10	0.46	8.62	0.015	0.08
GPT-2	0.72	8.53	0.020	0.11	0.72	8.01	0.022	0.09
Candela	0.69	19.31	0.072	0.10	0.70	22.22	0.022	0.09
GPT-2$_{KN}$	0.71	**6.85**	0.201	0.19	0.75	**6.33**	0.041	0.19
GPT-2$_{KN,MT}$	**0.85**	11.55	0.066	0.12	**0.86**	10.38	0.022	0.11
XNLG	0.83	6.94	**0.256**	**0.33**	0.84	8.15	**0.291**	**0.35**

Table 5: Results of CN generation with gold knowledge in-target and cross-target test sets.

in-target) scenario, we also explore the effect of knowledge infusion on cross-domain (i.e. cross-target) CN generation under zero-shot setting. We hypothesize that having a system trained to make use of substandard silver knowledge to generate proper CNs for a given context, could be robust to cross-domain zero-shot conditions. Therefore, we organized a data collection session with an expert operator in writing CNs. In this session, 50 islamophobic HSs randomly sampled from CONAN and 144 new cross-target HSs (covering misogyny, antisemitism, racism, and homophobia) are provided along with the knowledge retrieved by $Q_{hs \cup cn}$ queries. The expert is tasked with composing a suitable CN using the corresponding knowledge as much as possible. Thus, we could obtain a gold test set[10]. in which the input knowledge can certainly be found in the CNs.

We tested all models with gold knowledge in-domain and cross-domain test cases. Results are given in Table 5. For in-domain scenario, as we have anticipated, knowledge grounded models yield better performances in B-2 and R-L in comparison to the silver knowledge test setting. Especially with the striking jump in the performance of GPT-2$_{KN}$, we can confirm the proper infusion of the given knowledge to the generated CNs. As for cross-domain tests, GPT-2$_{KN}$ still yields better performance than baselines while the performance for all models (except for XNLG) drops due to unseen events during training. All GPT-2 variations present better diversity performances on the cross-domain setting as compared to both in-domain and silver-knowledge settings. Regardless of domains, XNLG yields fallaciously high scores due to its extensive copying. A cross-target generation from the models can be seen in Table 7. More examples in-/cross-domain generations from all the models are provided in Appendix A.6.

Human evaluation. We further resort to human evaluation to assess the final generation quality of

[10] We release the gold test set at https://github.com/marcoguerini/CONAN.

each model. For this reason we perform human evaluation of generation using gold knowledge, to rule out the effect of possible noise in the knowledge that may result from the retrieval process.

Our models were evaluated by 3 expert operators from the NGO Stop Hate UK. The annotators are already experienced, and specifically trained, in reading hateful content and writing CNs for online hate countering[11]. The annotators are instructed to assess all generated pairs in gold knowledge test sets in terms of *suitableness* to the HS, *informativeness*, and *intra-coherence* of CN regardless of HS. Each score is on a scale of 1 (the least) to 5 (the most). To avoid possible bias and hints towards models, we normalized the pairs (e.g., lowercase and space between words and punctuation) and divided them into 3 partitions of randomized files for experts (See Appendix A.5 for annotation instruction). Each expert was given 388 pairs, resulting in a total of 1164 pairs for evaluation. To avoid excessive workload annotators were allowed to complete the task over multiple sessions at their preference.

Results are reported in Table 6. We also computed Kendall's Tau-b (Kendall, 1938) to measure the annotators' agreement towards the model ranking for each aspect. The high correlations indicate a strong concordance among the annotators (threshold tau-b > 0.35). Regardless of domains, annotators consider XNLG generations as the most informative and GPT-2$_{KN}$ generations as the most suitable. TRF yields a reasonable suitableness and coherence since it tends to memorize the training CNs, almost behaving like a retrieval system on human responses. However, such behavior can be fatal in cross-domain settings. Candela fails to generate suitable cross-domain CNs despite preserving the intra-CN coherence. While GPT-2 and GPT-2$_{KN}$ generations are found almost equally coherent, the lower suitableness and informativeness of GPT-2 output (2.26 and 1.92) for cross-domain as compared to GPT-2$_{KN}$ (2.51 and 2.29) encourages the grounding CNs in knowledge.

	in-domain			cross-domain		
	suit.	info.	cohe.	suit.	info.	cohe.
TRF	2.65	2.25	3.39	1.47	2.09	3.45
GPT-2	2.67	2.16	4.10	2.26	1.92	**4.24**
Candela	2.41	2.25	3.14	1.42	2.09	3.40
GPT-2$_{KN}$	**3.02**	2.35	**4.33**	**2.51**	2.29	4.21
GPT-2$_{KN,MT}$	1.76	1.65	3.73	2.03	1.76	3.88
XNLG	1.43	**3.88**	2.12	1.88	**4.10**	2.79
Kendall's tau-b	0.82	0.69	0.82	0.51	0.91	0.73

Table 6: Human evaluation results of CN generation.

7 Discussion

Our findings suggest that a large pre-trained LM with knowledge injection is preferred to alleviate the demand for gold data and improves in-/cross-domain generations. GPT-2$_{KN}$ outperforming GPT-2, which becomes more clear with every increase in the quality of provided KN (i.e., from silver Q_{gen} to silver $Q_{hs \cup gen}$ to gold in-domain), indicates that our model trained on silver knowledge learns to grasp relevant KN to the given HS instead of random copying. As for GPT-2$_{KN,MT}$, since the model is trained to distinguish CN from HS, certain typical counter responses are prone to appear. Through a manual analysis, we confirmed that its responses tend to be simpler (e.g., negation: "*terrorism is **not** a religion.*", "*It's **not** true...*") than the responses of other GPT-2 variant models. Future work can explore alternative multitask learning frameworks to improve the generation.

From a practical perspective, we can conclude that fine-tuning a pretrained language model for conditional generation, utilizing silver-knowledge sentences retrieved through a query generation model is the most promising approach for knowledge-bound CN generation. In fact, (i) among various models we tried, GPT-2 with knowledge infusion is shown to produce the most informative and suitable output. (ii) At deployment time, using input context together with generated keyphrases becomes prominent with more novel and diverse CNs and (iii) for new hate-targets in a zero-shot scenario, the generation model can produce coherent and suitable CNs without retraining if a reasonable set of input knowledge is provided. Moreover, this approach can be applied to other scenarios, such as dialogue response generation, since it does not require any extra data collection or annotation.

8 Conclusion

Online hate speech intervention is a challenging problem and research on counter narrative generation is still in its infancy. In this work, we have proposed methods for improving counter narrative generation to fight hatred online, incorporating external knowledge retrieved through extracted and generated keyphrases. To this end, a dataset of hate-speech/counter-narrative pairs was augmented with relevant knowledge to train systems that are able to

[11] The compensation for annotation work met with EU regulations.

HS: Jews are different from non-Jews not because of their beliefs, but because of their physical nature. “Jewishness” and malevolence is inherited through blood.
Human: Today one of the biggest concerns is the reinvention of clinical concepts of race, based on inaccurate generalizations. The notion that the superiority of the white race was part of the natural order was deviously reinforced by the rise of modern ”scientific” racism in the late 19th century. Despite the strong objections of many of our contemporaries, this idea of a ”jewish race” is still going around.
GPT-2: Muslims are not different from non-Muslims because of their faith.
GPT-2$_{KN}$: I think that the notion of race is a generalization based on ignorance.
GPT-2$_{KN,MT}$: All that is wrong is that we cannot accept that violence is necessary.
XNLG: today , one of the biggest concerns is the reinvention of clinical concepts of race , based on inaccurate generalizations about the prejudice to certain physical characteristics and civilizations . despite the strong objections of many contemporary contemporaries , the notion of fixed ” race ” - packages of physical and behavioral characteristics

Table 7: Samples of cross-domain generation.

produce suitable and informative arguments. Our experiments on in-/cross- domain generation indicate that the generated responses can meet these desiderata. As future work we plan to test other query generation approaches (e.g., exploit human-crafted queries in an interactive setup or the use of pre-trained LMs) to improve knowledge selection and to test other architectures for the final generation step.

Acknowledgments

This work was partly supported by the HATEMETER project within the EU Rights, Equality and Citizenship Programme 2014-2020. We are deeply grateful to Stop Hate UK and its volunteers for their help and effort in evaluating the output of our systems.

Ethical Considerations

While we believe that counter narratives are a better tool than content moderation in fighting hate speech (e.g. they do not hinder freedom of speech), still the automatic generation of CNs should be taken with care. Since this work aims at presenting a methodology for knowledge-bound counter narrative production through neural approach, several ethical consequences should be considered.

First, neural models may still produce substandard counter narratives containing abusive language or negative content. To mitigate this issue, possible solutions include (1) integrating in the pipeline a classifier or a human reviewer for validation and possible post-editing (Tekiroğlu et al., 2020), (2) detoxification techniques for controllable generation methods (Gehman et al., 2020), and (3) discarding undesirable content from the corpora used for training (Raffel et al., 2020), even if the appropriate criteria for such purpose are still investigated.

Second, while our approach reduces the risks of content hallucination, an additional step, where the accuracy of the generated text is checked against the provided knowledge (Nie et al., 2019; Dušek and Kasner, 2020), would provide further robustness to the system.

Third, natural language generation models may still induce unintended social biases. This issue can be moderated by measuring/promoting fairness in models and data employed (Blodgett et al., 2020), and designing bias triggers (Sheng et al., 2020) or regularization methods (Bordia and Bowman, 2019; Corbett-Davies et al., 2017) for controllable bias.

To sum up, while some additional automated techniques may help in maintaining generation quality, human evaluation should always be considered as the foremost solution, at least for delicate tasks such as ‘real’ hate countering on social media platforms. For this reason we advocate that generation systems should be used as a suggestion tool for NGO operators, to make their countering work more effective. In this way there is always a “human moderator” taking the final decision (Chung et al., 2019).

References

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: United States Holocaust Memorial Museum*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016a. Considerations for successful counterspeech. *Dangerous Speech Project*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016b. Counterspeech on twitter: A field study. *Dangerous Speech Project*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7570–7577. AAAI Press.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–365, Vancouver, Canada. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations*, New Orleans, LA.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Julian Ernst, Josephine B Schmitt, Diana Rieger, Ann Kristin Beier, Peter Vorderer, Gary Bente, and Hans-Joachim Roth. 2017. Hate beneath the counter speech? a qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, (10):1–49.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Maurice George Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Dieu-Thu Le, Cam Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv:1812.02712*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. *International Conference on Learning Representations*.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the dirt: Extracting keyphrases from texts with kd. In *Proceedings of the Second Italian Conference on Computational Linguistics*, Trento, Italy.

Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Emanuele Pianta and Sara Tonelli. 2010. KX: A flexible system for keyphrase eXtraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 170–173, Uppsala, Sweden. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Unpublished.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication*.

Amrita Saha, Ghulam Ahmed Ansari, Abhishek Laddha, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2019. Complex program induction for querying knowledge bases in the absence of gold programs. *Transactions of the Association for Computational Linguistics*, 7:185–200.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ICA Annual Conference*, pages 1–23, Fukuoka, Japan.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Tanya Silverman, Christopher J Stewart, Jonathan Birdwell, and Zahed Amanullah. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue*, pages 1–54.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv:1908.09203*.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1177–1190, Online. Association for Computational Linguistics.

Henry Tuck and Tanya Silverman. 2016. *The counter-narrative handbook*. Institute for Strategic Dialogue.

Peter D Turney. 2000. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4):303–336.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Vincent Wang. 2018. Sketching a Chinese writer's vocabulary profile in English: the case of Ha Jin. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2018. Transfertransfo: A transfer learning approach for neural network based conversational agents. *NeurIPS 2018 workshop on Conversational AI: "Today's Practice and Tomorrow's Potential"*.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62, Vancouver, BC, Canada. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

A Appendices

A.1 Analysis on Keyphrase Extraction Configurations

We conducted a preliminary manual analysis to investigate the effects of various keyphrase extraction configurations. We randomly sampled 48 hate speech and counter-narrative pairs from CONAN dataset and extracted the keyphrases. Then, we retrieved the KN (see Section 5.3) with the queries Q_{hs} and $Q_{hs \cup cn}$. On the other hand, we also wanted to inspect the condition with the keyphrases only from CN, i.e., Q_{cn}. For each sample and condition, annotators have assigned a score for relevance to the hate speech in the scale of 1 to 5; 1 meaning no-relevance, and 5 perfect relevance. As a result, we have noticed that Q_{cn} is the worst condition, i.e., non-optimal, having an average score of 2.30. The analysis shows that it causes the loss of context related to HS, bringing information mostly from whole another topic. For instance, especially when CNs are rather generic, often, no lexical hint can be found related to the topic of Islamophobia (e.g., "*Do you have any proof?*"). Indeed, Q_{hs} provides an apparently better average score (3.46) since it provides a better context to search for. However, as expected, the best score (3.77) has been obtained through $Q_{hs \cup cn}$, i.e, optimal, verifying our hypothesis of utilizing both HS and CN keyphrases for training.

A.2 Preliminary Analysis on Decoding Methods for XNLG

To find a suitable decoding method for our task, we generated CNs with 3 candidate settings: beam search with a beam-width of 3 and top-k sampling with a k value of 8 and 10. For each setting we utilized KN retrieved with both non-optimal (Q_{cn}) and optimal ($Q_{hs \cup cn}$) queries. Then we sampled 120 HS-CN pairs and served them to three experts in CN writing for evaluating the generation on a scale of 1 (the worst) to 5 (the best) in terms of suitableness and informativeness. Suitableness measures if the generated CN is relevant to the HS and informativeness evaluates the amount of information (e.g. statistics and facts) enclosed in the CN.

The results reported in Table 8 reveal a clear difference between beam search and top-k sampling regardless of KN being optimal or non-optimal. In a manual investigation, we observed that the generation using both beam search and top-k sampling generally can copy some pieces of information from the given KN, while top-k seems to replace part of the text with slightly relevant and uncommon words. Hence, copying the right knowledge pieces through the decoding strategy is a key factor instead of diverging from the knowledge solely for the sake of lexical diversity. Therefore, based on the results, we adopt beam search with a beam-width of 3, which is shown to be the most suitable and informative, for decoding method in our experiments.

Decoding methods	Suit.	Info.
Non-optimal knowl.		
Beam-3	1.950	2.325
Topk-8	1.275	1.775
Topk-10	1.625	2.100
Optimal knowl.		
Beam-3	2.325	2.450
Topk-8	1.975	2.175
Topk-10	2.050	2.325

Table 8: Human evaluation on CN generation using various decoding methods.

A.3 CN Generation with Q_{hs}

In this section we report the CN generation results of our knowledge-bound models trained and tested with Q_{hs}. We applied the same hyperparameter configurations as the models trained with $Q_{hs \cup cn}$ described in Section 6.1.

The results are given in Table 9. In contrast to the baselines (i.e., models without knowledge and Candela), all models obtained higher novelty with Q_{hs}. The repetition rate, on the other hand, is not improved since the models exploit the same knowledge for multiple test samples due to the repeated HSs with different CNs in the test set.

We also observed that for GPT-2_{KN} and GPT-$2_{KN,MT}$ the generation with Q_{hs} is more repetitive and less novel compared to the generation applying queries $Q_{hs \cup gen}$ (as shown in Table 4). This result demonstrates the viability and necessity of using generated queries, as potential CN prompts, along with HS context.

A.4 Unique HS Test Set Analysis

Concerning that one HS can be paired with different CNs in the test, we further conducted an evaluation on a unique set by keeping each unique HS and one randomly selected CN among its CNs. The unique HS set lets us perform a fairer comparison among query configurations especially for Q_{hs}

							KN overlap (ngram)		
Models	Nov.	RR	B-2	R-L	#Word	#Sent.	1	2	3
GPT-2$_{KN}$	0.700	8.66	0.082	0.098	14.87	1.32	0.235	0.019	0.003
GPT-2$_{KN,MT}$	0.730	11.92	0.084	0.090	13.25	1.13	0.233	0.025	0.007
XNLG	0.824	16.46	0.073	0.084	55.51	3.71	0.841	0.650	0.558

Table 9: Results of CN generation applying Q_{hs}.

Models	Nov. (W/U)	RR (W/U)
without knowledge		
TRF	0.467/0.457	7.72/6.97
GPT-2	0.688/0.675	9.04/7.79
Train$_{cn}$	-	3.91/3.31
with knowledge		
Candela (Q_{hs})	0.692/0.697	21.87/21.97
GPT-2$_{KN}$		
w/ Q_{hs}	0.723/0.719	8.13/7.97
w/ Q_{gen}	0.728/0.720	7.48/6.34
w/ $Q_{hs \cup gen}$	0.735/0.740	6.30/5.81
w/ $Q_{hs \cup cn}$	0.727/0.731	7.17/6.16
GPT-2$_{KN,MT}$		
w/ Q_{hs}	0.744/0.748	11.69/10.24
w/ Q_{gen}	0.731/0.750	10.37/9.25
w/ $Q_{hs \cup gen}$	0.747/0.747	7.59/8.41
w/ $Q_{hs \cup cn}$	0.731/0.728	9.56/11.08
XNLG		
w/ Q_{hs}	**0.824/0.828**	14.42/6.22
w/ Q_{gen}	0.819/0.821	6.88/**4.03**
w/ $Q_{hs \cup gen}$	0.812/0.817	6.98/5.19
w/ $Q_{hs \cup cn}$	0.819/0.816	**5.69**/4.64

Table 10: Results of unique HS test set (U) for novelty and diversity. For comparison, we keep the results of whole test set (W).

with models employing beam search (i.e. XNLG and Candela). The results are given in Table 10.

For XNLG and GPT-2$_{KN,MT}$, we observed an increase in the novelty and diversity with Q_{hs} and Q_{gen} on the unique HS set over the whole test set. As for GPT-2$_{KN}$, while diversity improves for all query configurations, we did not observe an increase in novelty through Q_{hs}. For Candela, while novelty also increases, the diversity does not improve.

A.5 Human Evaluation Instruction

Three experts are tasked with CN generation evaluation. We briefed the annotators the aim of the evaluation, and then we explained common and appropriate CNs along with 10 examples of HS-CN pairs. The annotators are instructed to evaluate the generation with intuition. The generation is evaluated in terms of 3 aspects: (i) **suitableness:** How suitable the counter narrative is under evaluation for its hate speech, i.e. coherence and relevance of a counter narrative with regard to the hate speech. (ii) **informativeness:** How informative/specific the counter narrative is as a response for its hate speech. (iii) **intra-coherence:** How coherent the counter narrative is under evaluation as a response, i.e. internal coherence of a counter narrative regardless of its hate speech. In Table 11 we give the annotation examples with the score explanations provided to the annotators.

HS: Muslims do not want to assimilate. They demand that everyone shows respect for Islam, but they don't respect our laws. One day they will oppress us with the Sharia and polygamy.
CN: Do you really believe that they are a problem? I have a job. **evaluation**: suit.: 3, info.: 1, cohe.: 1 **Note**: The first sentence is relevant but not the second one. There is no coherence between 2 sentences.
CN: As far as they do not force us to follow sharia laws, we should not be concerned with them following their laws. **evaluation**: suit.: 5, info.: 3, cohe.: 5 **Note**: The CN is an appropriate and coherent response, but not providing a substantial amount of information or facts.
CN: Muslims should not be forced to assimilate, since it is not right and no one wants that. And polygamy is illegal and forbidden in UK and Muslims actually respect this ban. **evaluation**: suit.: 5, info.: 5, cohe.: 5 **Note**: The whole CN is perfectly suitable for the given HS. It is highly informative with the facts and internally coherent.

Table 11: Annotation examples with score explanations.

A.6 Generation Examples

HS: Jews are Christ Killers.
Human: While the Catholic Church has for decades taught that Jews weren't collectively responsible, the argument laid out by the Pope Benedict XVI was a landmark statement. In his book, he not only confirms church teaching refuting the decide charge, he also explains biblically and theologically why there is no basis in Scripture for the argument that the Jewish people as a whole were responsible for Jesus' death.
TRF: Why do you consider muslims like beasts? Did they do something wrong? If you take into consideration the crimes committed by muslims, why do you distinguish the muslim criminals from the non muslim? They are all criminals and all of them are punished.
GPT-2: I know many Muslims who are peaceful and respectful and do not kill.
Candela: Do you have proof? Where did you get this information from? I do not see the link between islamism and the news of the muslim community.
GPT-2$_{KN}$: No, the crucifixion of Jesus Christ is not a source for anti-Semitism. The Jewish people were not responsible for the death of Jesus.
GPT-2$_{KN,MT}$: On the basis of christianity, as christianity is not a religion, jews are not responsible for their own death.
XNLG: In ”jesus of nazareth-part ii” excerpts the church's teaching that refuted the deicide charge of the jewish people, as well as theologically explains one of the most controversial issues in christianity.
KN: Pope Benedict XVI has made a sweeping exoneration of the Jewish people for the death of Jesus Christ. While the Catholic Church has for five decades taught that Jews weren't collectively responsible, Jewish scholars said... would help fight anti-Semitism today. In ”Jesus of Nazareth-Part II” excerpts released Wednesday, Benedict explains biblically and theologically why there is no basis in Scripture for the argument that the Jewish people as a whole were responsible for Jesus' death.

Table 12: Sample cross-domain CN generation from different models.

HS: Why gays do not try to heal?
Human: Contrary to some insinuations, sexual orientation is not like a disease.
TRF: Type imam boussena and we ll see who the scum is.
GPT-2: So they're the ones who are trying to make us hate them.
Candela: You see more and more attacks on muslims as a result of the controversies? Do you really want to believe in a 'caliphate'?
GPT-2$_{KN}$: We know that homosexuality is not a disease.
GPT-2$_{KN,MT}$: The genetics of homosexuality have been studied, but it seems to me that the human race is not suited to do that.
XNLG: And , contrary to the insinuation , sexual orientation is not like a disease . more recent studies have identified other potential links between homosexuality and genetics , and scientists have offered explanations on how same sex attraction can advance procreation , suggesting that the same twin is more likely to be gay , but finding the biological origins of the same
KN: Contrary to Perry's insinuation, sexual orientation is not like a disease. More recent studies have identified other potential links between homosexuality and genetics, and scientists have offered explanations on how same-sex attraction could advance procreation. Studies of identical twins, who share the same genes, have shown that, if one twin is gay, the other is more likely to be gay, suggesting sexual orientation is influenced by genetics. Scientific studies have demonstrated that biology plays a role in one's sexual orientation.

Table 13: Samples of cross-domain CN generation from different models.

HS: It seems that I cannot define Islam as a cult which promote wars and conquers host nations and legitimate this by its scriptures. Can I define them as narrow-minded, prejudiced, sexist and male chauvinist?
Human: Everyone has the right of free speech, while it does not mean anyone can use it to hurt people. Plus, based on what you critisize Islam? I recommend you to read something about true Islam and Muslims. You would find that they are no difference from us.
TRF: Islam is all about the religions and tolerance insert quote from koran. Migrants is not evil?
GPT-2: It seems to me that this is the antithesis of what Islam teaches.
KN Q_{hs}: Hakim al-Mutairi, secretary-general of the emirate's Umma (Islamic Nation) party, urged the pope to apologise immediately "to the Muslim world for his calumnies against the Prophet Muhammad and Islam".... Instead of bringing Islam and Christianity closer, he is straining relations between the two religions". In the present political atmosphere such views can be exploited by those who are trying to malign Muslims and Islam. **Candela:** I have no accountability. I have no problem with those who claim to belong to, but i do not have the right to practice whatever religion. **GPT-2$_{KN}$:** The Qur'an clearly teaches peace, love and tolerance towards everyone, including the non-Muslims. The only thing that Islam does not have is a message of conquest and war. **GPT-2$_{KN,MT}$:** Islam is a religion founded on the principle that you must believe in the good will of all people. **XNLG:** Extremist muslim national party secretary, hamid al mutairi, urged the pope to define immediately to the muslim world for his calumies against the prophet muhammad and islam. Pope benedict xvi was full of hostility and hatred against muslims and islam. In the present political atmosphere such views can be exploited by those who bring
KN $Q_{hs\cup gen}$: Islam, as practiced by the vast majority of people, is a peaceful religion, a religion that respects others. "The large majority of Muslims tinker," said Franck Fregosi, a sociologist who has written extensively on Islam in Europe.... **GPT-2$_{KN}$:** Hi, I do not understand the problem you have. The vast majority of Muslims do not advocate wars and conquer the world. They practice a peaceful religion that respects others. **GPT-2$_{KN,MT}$:** It seems that i have just come across the iota of muslims who have been insulted, persecuted, killed, raped, harassed, murdered, killed by the muslims in europe. **XNLG:** It seems that islam, as practiced by the vast majority of people, is a peaceful religion, a religion that respects others. The large majority of muslims tinker, as franck fregosi, who has written extensively on islam in europe. It seems that it does not define islam as an inclusion and respect for all of the religion
KN $Q_{hs\cup cn}$: No doubt, it is the legitimate right of Muslims to build a community center near Ground Zero. So please stop conflating criticism of Islam with discrimination of Muslims. As a nation of free people, we can never defeat Islam but we can make war against terrorists and governments that support them. a "moderate Islamic countries" when all unanimously reject the Universal Declaration of Human Rights and support The Cairo Declaration of Human Rights in Islam. **GPT-2$_{KN}$:** The Islamic countries do not promote wars and conquer host nations. Islam is not a cult but a peaceful religion which respects all peoples. **GPT-2$_{KN,MT}$:** The vast majority of men in the world are not able to articulate the spiritual values of islam. **XNLG:** It seems that it is the legitimate right of muslims to build a community center near zero ground, so please confine criticism of islam with discrimination against muslims. As a nation of free people, we can never defeat islam, but we can make war against terrorists and governments that support them. My comment gives a lot of information about what islam and muslims

Table 14: Sample of in-domain generation.