Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation

Jaione Bengoetxea[1], **Yi-Ling Chung**[2], **Marco Guerini**[3], **Rodrigo Agerri**[1]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU,
[2]The Alan Turing Institute,
[3]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
jaione.bengoetxea@ehu.eus, ychung@turing.ac.uk, guerini@fbk.eu, rodrigo.agerri@ehu.eus

Abstract

Counter Narratives (CNs) are non-negative textual responses to Hate Speech (HS) aiming at defusing online hatred and mitigating its spreading across media. Despite the recent increase in HS content posted online, research on automatic CN generation has been relatively scarce and predominantly focused on English. In this paper, we present CONAN-EUS, a new Basque and Spanish dataset for CN generation developed by means of Machine Translation (MT) and professional post-edition. Being a parallel corpus, also with respect to the original English CONAN, it allows to perform novel research on multilingual and crosslingual automatic generation of CNs. Our experiments on CN generation with mT5, a multilingual encoder-decoder model, show that generation greatly benefits from training on post-edited data, as opposed to relying on silver MT data only. These results are confirmed by their correlation with a qualitative manual evaluation, demonstrating that manually revised training data remains crucial for the quality of the generated CNs. Furthermore, multilingual data augmentation improves results over monolingual settings for structurally similar languages such as English and Spanish, while being detrimental for Basque, a language isolate. Similar findings occur in zero-shot crosslingual evaluations, where model transfer (fine-tuning in English and generating in a different target language) outperforms fine-tuning mT5 on machine translated data for Spanish but not for Basque. This provides an interesting insight into the asymmetry in the multilinguality of generative models, a challenging topic which is still open to research.

Content Warning: This paper contains examples of offensive language that do not reflect the authors' views.

Keywords: Counter Narratives, Hate Speech, Multilinguality, Text Generation

1. Introduction

In the last few years, and partially due to the anonymity of citizens when interacting in the online world, hate speech (HS) has become an ever-increasing media presence, to the point of being almost normalized. Davidson et al. (2017) defined HS as "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group."

Currently, online sites and social media platforms are constantly updating their policies to fight the ever evolving online hate, with the majority taking a block-and-delete approach. In order to facilitate the workload that these policies could produce, automatic detection of HS has become a very active research field, including the development of HS datasets (Basile et al., 2019; Kolhatkar et al., 2020) and machine learning and deep learning techniques (Nobata et al., 2016; Davidson et al., 2017; Faris et al., 2020).

It has been argued that while such methods based on content moderation can immediately reduce the amount of HS, it draws limits on free speech (Schieb and Preuss, 2016) and may not be effective at challenging HS in the long term. For example, it does not address the root causes that lead to the HS or attempts to change the outlook of people who produce the HS. In this context, Counter Narratives (CNs) have been proposed as an effective approach to tackle and mitigate the spread of HS (Benesch, 2014; Schieb and Preuss, 2016). Counter Narratives are non-aggressive responses to a hateful comment that includes non-negative factual-based argumentative feedback (Benesch, 2014; Schieb and Preuss, 2016). An example of an HS-CN pair can be observed below (Chung et al., 2019):

> **Hate Speech** Islamic are criminals: they rape, enslave and murder people. Islam is more a worship than a religion and we do not have anything to share with them.

> **Counter Narrative** The myth that Muslims are dangerous and violent is a product of our vilifying media. Don't believe everything you read.

While recent interest in automatic approaches to CN studies has grown considerably, including studies on data curation (Chung et al., 2019; Fanton et al., 2021), detection (Chung et al., 2021a; Mathew et al., 2018), and generation (Tekiroglu et al., 2020; Chung et al., 2021b; Zhu and Bhat, 2021; Tekiroğlu et al., 2022), experimental work on automatic CN generation has been predominantly carried out for English. This is due to the lack of

both non-English manually curated training data (Chung et al., 2020) and of generative language models pre-trained for a large number of languages (Xue et al., 2021; Lin et al., 2022).

In this paper we address the first shortcoming by presenting CONAN-EUS, a new parallel Basque and Spanish dataset for CN generation developed by taking the full English CONAN as a starting point and applying Machine Translation (MT) and professional manual post-edition. The corpus consists of 6654 machine translated HS-CN pairs and 6654 gold-standard human-curated HS-CN pairs (per language) parallel to the original English CONAN (Chung et al., 2019) accounting for a total of 26616 entries. This new resource allows to perform novel research investigating the impact of *data-transfer* (train on translated data for the target language) and *model-transfer* (train in a given source language, e.g. English, and generate in a different target language using some multilingual language model) on the task of CN generation.

Thus, empirical experimentation in all three languages, Basque, English, and Spanish with the multilingual mT5 encoder-decoder model (Xue et al., 2021) demonstrates that CN generation is substantially better when the model is fine-tuned on post-edited data. Most importantly, the results of a manual qualitative evaluation correlate with this result, demonstrating that manually revised training data still makes a difference in generating better-quality CNs.

Furthermore, *multilingual data augmentation* (fine-tuning mT5 on all three languages at the same time and generating CNs for each target language in turn) improves results over the monolingual setting (fine-tuning and generating in the same language) for structurally similar languages such as English and Spanish. However, multilingual data augmentation is detrimental for Basque, a language isolate and markedly different from English and Spanish.

A similar phenomenon was found in the zero-shot crosslingual evaluations, where applying mT5 in a *model transfer* approach outperforms the translate-train *data-transfer* approach for Spanish, while the opposite is true for Basque. We hypothesize that this might be related to the asymmetry in the multilinguality of both decoder and encoder models, a challenging topic which is still open to research (Artetxe et al., 2020; García-Ferrero et al., 2022; Lin et al., 2022; García-Ferrero et al., 2022). We believe that our work opens new avenues for research on CN generation and crosslingual transfer with the aim of not only improving current techniques for CN generation, but also to mitigate the lack of manually curated training data for many languages. Data, code and fine-tuned models are publicly available[1].

---

2. Related Work

Research on CN has become a topic of interest in recent years, so in this section we will mostly review previous work on CN generation methods and datasets from a multilingual perspective.

Counter Narrative datasets Qian et al. (2019) used two different data sources to build a dataset consisting of online HS instances collected through a keyword search. CNs were then manually written via crowdsourcing.

Another popular dataset is CONAN (COunter NArratives through Nichesourcing) (Chung et al., 2019). CONAN includes HS-CN pairs on Islamophobia, with the aim of providing a non-ephemeral, expert-based, multilingual corpus of HS-CN pairs in English, French and Italian. Nichesourcing was chosen to prioritize the quality of the generated data - a process in which complex computational tasks are completed by experts in the specific field of the task (Boer et al., 2012). CONAN was created through separate data collection sessions, which involved the help of more than 100 NGO expert operators on countering hate online which involved more than 500 person-hours of data collection (Chung et al., 2019).

While crowdsourcing processes facilitate the annotation of large amounts of data, for this kind of tasks the result can often be simple and stereotypical (Tekiroglu et al., 2020). In contrast, even though nichesourcing is a more laborious process, the generated CNs appear to be more diverse and less stereotypical, as they are created by experts in the field, thus suggesting better end results in terms of quality (Chung et al., 2019).

As collecting data through experts is time-consuming and costly, two simple data augmentation processes were applied to CONAN by three non-expert annotators, which required each of them working for around 200 hours. First, two paraphrases of each original CN for each of the languages were created by non-expert workers. Second, the original French and Italian HS-CN pairs were professionally translated into English. In doing so, a final English corpus of 6654 HS-CN pairs was created. Details about the CONAN dataset are provided by Table 1.

In this work, we build CONAN-EUS by (i) automatically translating the 6654 English HS-CN pairs into Basque and Spanish and, (ii) giving the automatic translations to professional translators for post-editing. Manual post-edition was considered because, although being costly, it is still order of magnitudes lower than building a similar resource from scratch. Furthermore, an additional benefit is that, by generating parallel data, we aim to facilitate crosslingual research on CN generation. This

	En	Fr	It	Total
Original	1288	1719	1071	4079
Augmented	2576	3438	2142	8159
Translated	2790	-	-	2790
Total	6654	5157	3257	15025

Table 1: Number of HS-CN pairs in the original CONAN dataset. Augmented: two paraphrases of the original data. Translated: manual translation of the original Italian and French into English.

data collection process is detailed in Section 3. To our knowledge, no previous work on Basque CN datasets is available.

Counter Narrative generation models Qian et al. (2019) experimented with three encoder-decoder generation models establishing a baseline that could be used for future generation studies. Moreover, Pranesh et al. (2021) found out that testing various models for CN generation resulted in a lack of correlation between the quantitative and qualitative manual evaluations. These dissonances between automatic and manual evaluations were also previously observed (Qian et al., 2019).

The large majority of previous work has noted a lack of diversity and relevance with respect to the HS in the generated CNs (Fanton et al., 2021; Qian et al., 2019). In order to address these issues, Zhu and Bhat (2021) proposed a three-module pipeline named "Generate, Prune, Select" (GPS). This involved the generation of phrases, the pruning of ungrammatical sentences, and the selection of the most relevant responses to the given HS. Their findings showed that GPS improved both diversity and relevance aspects when compared to several baseline models.

Other works focused on the factuality and veracity of CNs, introducing knowledge-grounding techniques to improve the suitability and informativeness of CN generation (Chung et al., 2021b). More recently, Tekiroğlu et al. (2022) made an extensive comparison of pre-trained language models, finding that decoder models such as GPT-2 and DialoGPT performed best in terms of specificity and novelty. Additionally, they proposed a pipeline where an automatic post-edition step could be added in order to refine the generated CN. We believe that automatic post-edition techniques will benefit from the availability of parallel corpora such as CONAN-EUS that introduces machine translated HS-CN pairs together with their human-curated counterparts, as in the original English CONAN (Chung et al., 2019).

Counter Narrative generation in other languages As far as we know, only two works have been published on CN generation for languages different from English. First, Chung et al. (2020) presented a number of experiments for CN generation in Italian aiming to study the effects of using silver MT data as opposed to manually generated. Their results show that using MT data as a form of data augmentation, combined with gold data, helped improve performance. This method is particularly interesting to cheaply obtain training data, given that collecting gold data is often unfeasible in terms of manual cost and time. Second, Vallecillo-Rodríguez et al. (2023) translated CONAN-KN dataset into Spanish(Chung et al., 2021b), composed of 238 HS-CN pairs and, probably due to the tiny size of their data, focused on experiments using few-shot and prompting methods.

More recently, Möhle et al. (2023) addresses the creation of CN data for non-English languages using the example of German. The increasing focus on data collection in diverse languages reflects a growing interest in CN generation in languages beyond English.

In contrast to prior work, our newly built CONAN-EUS provides the full 6654 HS-CN pairs from CONAN translated and then manually post-edited by professional translators into Basque and Spanish. Additionally, being a parallel dataset with the original English CONAN allows to undertake novel research on multilingual and crosslingual transfer techniques for the automatic generation of CNs.

3. Building CONAN-EUS

The 6654 English HS-CN pairs in the CONAN dataset were machine translated (MT) into Spanish and Basque, using the Google API[2]. Additionally, both Spanish and Basque MT datasets were post-edited by 3 native professional translators.

3.1. Basque Post-edition

Generally speaking, during the post-edition process of the Basque training set, three levels of post-editing could be identified. In the first case, no edits were made, as the MT sentence was correct. Second, minimal changes were needed (one word, termination or punctuation). For instance, the noun 'islam' was often translated as 'islam', but in the majority of the contexts in Basque this word needs the post-position '-a': 'islama'. And third, cases where more substantial changes were made, namely, whole expressions, word order, or the addition/deletion of whole sentences or clauses.

The most common error types were grammatical errors (verb tenses or conjugations, incorrect pro-

---

[2] https://pypi.org/project/google-trans-new/

		Spanish		Basque	
	Unique	Post-ed.	%	Post-ed.	%
HS	523	114	21.75	267	51.05
CN	4041	630	15.59	3605	89.21

Table 2: Post-edition statistics.

nouns), as we can see in Example 1[3], where the automatic translation used the verb in the present tense instead of the past.

Example 1

OG *Then why **did** we ask them to come in the first place*

MT *Orduan, zergatik eskatu **diegu** lehenbailehen etortzeko?*

PE *Orduan, zergatik eskatu **genien** etortzeko, lehenik eta behin?*

Moreover, lexical errors related with polysemy, or literal translation of expressions, were also detected. For instance, in Example 2 'free from conflict' is translated to 'doan gataztatik'. The sense of the word 'free' that has been translated in the MT is 'have it or use it without paying for it', which is in fact 'doan'. However, in this context the sense that we are looking for is more similar to 'not restrained to', which would be translated as 'libre'.

Example 2

OG *... only [some countries] are **free from conflict**.*

MT *... [herrialde batzuk] bakarrik daude **doan gatazkatik**.*

PE *... [herrialde batzuk] bakarrik daude **gatazkatik libre**.*

Some other errors noticed during post-edition were those related with the translation of acronyms or typos in the source English version.

Table 2 reports the statistics of the post-edition for both languages. It should be noted that during the nichesourcing of CONAN, NGO workers generated a number of CNs per HS (Chung et al., 2021a), which is why in the **Unique** column the numbers do not correspond to the 6654 HS-CN pairs contained in the dataset. The other two columns refer to the total number (**Post-ed.**) of single instances (HS or CN) and the percentage (**%**) that were post-edited.

With respect to Basque, it is quite clear that the percentage of post-edition is very high, which means that the quality of the machine translated

data was not that good. Furthermore, it can also be observed that CNs required more post-editing work than the HS, which could be due to the fact that CNs are in general longer and more complex.

3.2. Spanish Post-edition

During the Spanish post-edition process, some common errors were detected, which were also found in Basque. For instance, the automatic translation of acronyms is problematic in Spanish too. However, some other language-specific errors were found in Spanish, as for example gender concordance errors, caused by the lack of concordance in gender between some nouns with their reference noun. Thus, in Example 3 below, the two words in **bold** are written using the masculine gender mark 'o', but, as they refer to 'women' (mujeres) the correct form would be with an 'a', as in **verlas** or **orgullosas**.

Example 3

OG *Women of our culture that decide to become Islamic are so selfish. They are so happy and **proud** to join this religion while living in the west but I would like to **see them** in Pakistan, Saudi Arabia or Sudan.*

MT *Las mujeres de nuestra cultura que deciden volverse islámicas son muy egoístas. Están muy felices y **orgullosos** de unirse a esta religión mientras viven en el oeste, pero me gustaría **verlos** en Pakistán, Arabia Saudita o Sudán.*

If we look at the post-edition statistics of the Spanish portion of the dataset in Table 2, we can see that the post-edited percentage in both train and test set is considerably lower than those previously mentioned for Basque. Again, this indicates that the MT quality from English to Spanish was much better than for the English-Basque pairs.

4. Experimental Settings

Taking into account the automatic translation and manual post-edition to create CONAN-EUS, we have the following datasets available for experimentation:

---

[3]**OG:** original English CN, **MT:** machine translated version, **PE:** post-edited CN.

1. **Two MT datasets**, in Spanish and in Basque (henceforth, es-mt and eu-mt, respectively).

2. **Two post-edited datasets**, in Spanish and in Basque (henceforth, es-post and eu-post, respectively).

3. **The full CONAN English dataset**.

After generating three splits (4833 HS-CN pairs for train, 537 for validation and 1278 for test with no HS-CN pairs occurring across the splits), we took advantage of our new parallel CONAN-EUS dataset to devise three types of experiments: (i) **monolingual** in English, Spanish and Basque, consisting of fine-tuning mT5 and generating CNs for the same language, (ii) **multilingual**, with the aim of studying easy data augmentation by using all three languages for fine-tuning and, (iii) **zero-shot crosslingual** experiments, in order to evaluate the crosslingual model transfer capabilities of mT5 for CN generation.

In the monolingual setting we evaluated the performance of fine-tuning mT5 with: (i) automatically translated ($LANG-mt) data, and (ii) the post-edited ($LANG-post) data. Evaluation is always performed on post-edited data. With respect to the multilingual and crosslingual experiments, detailed in Table 3, we always used the post-edited data.

		Train	Test
Zero-shot	en2es	HS: en CN: en	HS: es CN: es
	en2eu	HS: en CN: en	HS: eu CN: eu
Multilingual	all2en all2es all2eu	en + es + eu	en es eu

Table 3: Experimental setup for multilingual and zero-shot crosslingual settings.

The models were trained using mT5 (Xue et al., 2021), a multilingual language model that uses a basic encoder-decoder Transformer architecture and was trained with the mC4 corpus, which contains text in 101 different languages, Basque among them. After some preliminary experimentation, we decided to use the mT5-base version provided by HuggingFace (Wolf et al., 2020). Grid search was applied for optimal hyperparameters, with the final configuration being: 50 epochs, 1e-3 learning rate, a batch size of 4 and 6 beam search. For reproducibility purposes, we also fixed a seed (42) for every experiment.

5. Experimental Results

We evaluated the performance of mT5 for the experimental settings specified in Section 4 using both automatic metrics and manual evaluation.

5.1. Automatic Evaluation

According to standard practice in previous work (Tekiroglu et al., 2020; Chung et al., 2020), the following automatic evaluation metrics were used: BLEU, which measures the n-gram correlation between the input and output text (Papineni et al., 2002); ROUGE-L, which finds the Longest Common Subsequence (LCS) between the reference and the candidate text and computes the precision and recall based on the LCS (Lin, 2004); and Repetition Rate (RR), which is computed by calculating the non-singleton n-grams that are repeated in the automatically generated text (Bertoldi et al., 2013).

5.1.1. Monolingual

Monolingual results for Basque and Spanish are reported in Table 4. The first noticeable result is that for both Basque and Spanish, post-edition of MT data helps to improve CN generation. Furthermore, while es-post obtains better scores in BLEU and ROUGE-L, results for Repetition Rate are better (the lower the better) with the original English data. This could imply that the high Spanish results in the n-gram overlapping metrics (BLEU and Rouge-L) might have been influenced by the repetitiveness of the output. Second, while eu-post is also better than eu-mt, its results are substantially lower than those of English or Spanish. Our hypothesis is that Basque might not be as well-represented in the common vocabulary of the mT5 model as English or Spanish.

Model	BLEU	Rouge-L	RR
en	9.81	16.82	5.73
es-mt	7.94	16.18	7.59
es-post	**11.23**	**18.99**	**6.32**
eu-mt	4.58	9.93	9.87
eu-post	**6.49**	**11.60**	**7.79**

Table 4: Monolingual results on post-edited test data. In **bold** best result per language and metric.

Consequently, these automatic evaluation results show that translate-train data-transfer approaches (fine-tuning on the translated set for the target language) rely on post-editing for optimal results, as fine-tuning with MT data does not reach the same level of performance.

5.1.2. Multilingual and Crosslingual

Results for multiligual and crosslingual experiments are presented in Table 5. On multilingual all2en and all2es results, it can be seen that fine-tuning on all three languages helps to improve results, especially for English. However, this is not the case for Basque (all2eu), which may suggest that multilingual data augmentation may work better for structurally similar languages, namely, English and Spanish.

With respect to the zero-shot crosslingual setting, the model transfer experiments (en2es and en2eu) obtain extremely different results. Crosslingual transfer to Spanish (en2es) performs reasonably well, outperforming the data-transfer translate-train approach (es-mt). However, model-transfer (en2eu) fails spectacularly in Basque, scoring well below eu-mt.

Model	BLEU	Rouge-L	RR
all2en	**10.79**	**17.19**	8.94
en2es	10.03	17.78	**5.15**
all2es	**11.36**	18.87	6.27
en2eu	2.81	7.26	12.40
all2eu	6.45	11.08	11.38
Monolingual baselines			
en	9.81	16.82	**5.73**
es-post	11.23	**18.99**	6.32
eu-post	**6.49**	**11.60**	**7.79**

Table 5: Multilingual and crosslingual results on post-edited test data. In **bold** best result per metric and language across all three settings: monolingual, zero-shot crosslingual and multilingual.

Summarizing, the data-transfer translate-train method (eu-mt) obtains better results than model transfer (en2eu) only for Basque. For Spanish it is the opposite. Furthermore, multilingual data augmentation beats the monolingual setting for English and Spanish, but it fails to do so in Basque. So far, the best strategy for Basque remains full post-edition of the training data, whereas for Spanish we could obtain competitive results using data-transfer (es-mt), or crosslingual transfer (en2es), without requiring to manually post-edit the full training data. We believe that the obtained results are indicative of the less-than-optimal performance of multilingual generative models such as mT5 for a less-resourced language like Basque, which is also structurally very different from English and Spanish.

5.2. Qualitative Evaluation

As several studies on CN generation have pointed out, automatic evaluation metrics such as the ones used in the previous section often fail to correlate well with human judgement (Pranesh et al., 2021; Qian et al., 2019). Thus, we undertook a manual qualitative evaluation and recruited two annotators who were native or proficient in both Basque and Spanish. Both are also experts in linguistics.

Five different aspects were considered, each of them capturing a different property pertaining to CNs: Relatedness, Specificity, Richness, Coherence and Grammaticality. In preparation for the manual evaluation, six models were chosen: es-mt, es-post, en2es, eu-mt, eu-post and en2eu. Regarding the data, 20 HS-CN pairs were randomly sampled from the output predictions of each model. The two annotators blindly evaluated all six models' outputs on the five criteria, on a scale from 1 to 5. The selection of the criteria was inspired by manual evaluations of Chung et al. (2019) and Chung et al. (2020).

1. **Relatedness**: It measures how related the CNs are with their corresponding HS, namely, whether the CN is relevant given the HS that it is responding to.

2. **Specificity**: It states whether the CN is rather generic or specific for the given HS it is responding to, thus replying to the question "can it be used for another completely different HS or not?"

3. **Richness**: In terms of language and vocabulary, it measures whether the CNs are simple or rather complex.

4. **Coherence**: It tells us whether the sentences make sense together, and if all ideas are clear and can be easily understood.

5. **Grammaticality**: It measures the grammatical correctness of the CNs.

Models	Cohen's Kappa
es-mt	0.8077
es-post	0.7742
en2es	**0.8407**
eu-mt	0.8504
eu-post	0.8929
en2eu	**0.9054**
Overall	0.8452

Table 6: Inter-annotator agreement.

Once all examples had been annotated, the average of each criterion and an overall score for each model were calculated, together with their respective standard deviations. We report the qualitative evaluation results in Table 7. Furthermore, Table

6 provides the inter-annotator agreement (IAA) between the two evaluators. Generally speaking, we can see that the IAA was rather high for all of the models evaluated, as well as for the overall evaluation. The high level of agreement suggests that the manual evaluation was reliable.

If we look at the results from the **monolingual settings**, it can be seen that the post-edited results are the best overall for both languages. Thus, comparing the MT with the post-edited results, it is possible to see that the largest differences occur in terms of Specificity and Relatedness. In other words, the models trained with MT data generate more repetitive and generic CNs than those fine-tuned with their post-edited versions.

The crosslingual results form a different picture for Spanish and Basque. The Spanish model trained in the crosslingual transfer setting (en2es) scores particularly well on Richness, Coherence and Grammaticality, and outperforms the data-transfer version es-mt for all five criteria. However, for Basque, the model-transfer setting (en2eu) is particularly low in terms of Relatedness and Specificity. Still, the overall scores of en2eu are better than those of eu-mt. In any case, it should be noted that for Basque the en2eu model-transfer approach slightly outscores the data-transfer method (eu-mt), the only result that does not correlate with the automatic evaluation discussed in the previous section.

Correlation Summarizing, the results for Spanish from both quantitative and qualitative evaluations correlate for every evaluation setting. This however, is not the case for Basque as the zero-shot crosslingual transfer (en2eu) is better according to the manual evaluation exercise but substantially worse in terms of the automatic metrics. In the next section we provide an error analysis to better understand this and other issues related with the automatic and the manual evaluation of CN generation.

6. Discussion

An error analysis was conducted in order to examine the quality of the generated CNs. In this section, we focused our attention on the Basque generated outputs. As previously mentioned, although eu-post performed surprisingly well in the manual evaluation, it still produces cases for which the scores in Relatedness and Specificity are rather low. An example of an unrelated and generic CN generated by eu-post is illustrated in Example 4. We can observe how the generated CN is not replying to the HS: while the HS is asking for the repatriation of Muslims, the CN is questioning why Muslims cannot build mosques in their country.

Example 4

HS *Musulmanak gure herrialdetik atera beharko lirateke inork ez dituelako onartzen.*

'Muslims should be taken away from our country because they aren't accepted by anyone.'

CN *Zergatik ez zaie baimendu behar musulmanei meskita hemen eraikitzea? Herrialde honek kristautasunaz gain beste erlijio batzuk praktikatzeko askatasuna ematean sinesten zuela uste nuen.*

'Why should not Muslims be allowed to build Mosques here? I thought this country believed in giving people the freedom to practice religions other than Christianity?'

A very similar phenomenon occurs when we look at the outputs generated by mT5 when fine-tuned in a zero-shot crosslingual setting (en2eu). Table 5 shows that en2eu obtained the worst overall scores across all models. However, en2eu performs considerably better in the manual evaluation, obtaining better overall results than eu-mt. After a closer analysis of the manual evaluation, it can be observed that although en2eu obtained the highest scores among the three Basque models for Richness and Grammaticality, its results for Relatedness and Specificity remain extremely low. Thus, even though mT5 generates grammatical CNs in an en2eu zero-shot crosslingual setting, they clearly fail to comply with the more domain-oriented human evaluation criteria (Relatedness and Specificity), as illustrated by Example 5.

Example 5

HS *Since 2000 all suicide bombings were committed by a Muslim and no one handle this fact that about global terrorism.*

CN *Musulman britainiarrek beste edozein talde erlijiosok baino gehiago ematen diote karitateari per capita [esteka MEND-eko estatistiketara edo/eta meskita bateko ongintzako dohaintza handien adibidera], beraz, kosta egiten zait ikustea funts horiek biltzen diren meskitekin zer arazoa dagoen.*

'British Muslims give more to charity, per capita, than all other religious groups [link to mend statistics and/or example of one large charity donation from a Mosque], so I am struggling to see the problem with Mosques where these funds are mostly raised.'

This example illustrates how the generated response, while perfectly fine in terms of grammaticality and richness, is neither related nor specific to the HS it is supposed to be responding to.

Finally, a note about the coherence of the generated CNs. Despite of the automatic evaluation, human evaluation shows that Coherence results

	Relatedness	Specificity	Richness	Coherence	Grammar	Overall
es-mt	2.30	2.50	3.33	3.58	3.78	3.10 ± 0.66
es-post	<u>3.61</u>	<u>3.31</u>	3.72	<u>4.25</u>	<u>4.33</u>	**3.84 ± 0.43**
en2es	3.20	2.90	<u>3.83</u>	4.00	4.08	3.60 ± 0.52
eu-mt	2.33	2.20	2.75	3.25	3.75	3.01 ± 0.65
eu-post	<u>3.13</u>	<u>3.08</u>	<u>3.15</u>	3.43	4.03	**3.85 ± 0.39**
en2eu	1.85	1.53	3.70	<u>3.98</u>	<u>4.43</u>	3.10 ± 1.31

Table 7: Qualitative results (average from annotators). <u>Underlined</u>: best per category and language; **Bold:** best overall per language.

are consistently high across all six evaluated settings. Therefore, our manual evaluation shows that the main problems of our approach to generate CNs are related with the Relatedness and Specificity criteria, and not so much with Coherence and Grammaticality.

In this sense, while we have followed the evaluation methodology from previous works using the original CONAN dataset, further research on the automatic evaluation of CN generation is needed as current automatic overlap-based metrics fail to capture application-specific features or aspects of the generated CNs.

7. Concluding Remarks

In this paper we have presented CONAN-EUS, a new parallel Basque and Spanish dataset for CN generation consisting of automatic translations and professional post-editions of the original English CONAN. The corpus consists of 6654 machine translated HS-CN pairs and 6654 gold-standard human-curated HS-CN pairs (per language) which makes it a unique resource to investigate CN generation from a multilingual and crosslingual perspective.

Experimental results show that CN generation is better when mT5 is fine-tuned on post-edited training data, rather than on the output of MT. Multilingual experiments (training on all three languages) show that crosslingual transfer works best for English and Spanish, which could be due to a number of reasons: (i) English and Spanish and structurally more similar than with respect to Basque and, (ii) the quota of shared strings in the mT5 vocabulary is larger for those languages than for Basque, a less-resourced language. Evaluation on zero-shot crosslingual settings produced similar results. While model-transfer outperforms datatransfer for Spanish, the opposite is true for Basque. Still, crosslingual transfer in generative approaches remains an open and complex research problem (Lin et al., 2022).

Our work underpins several opportunities for future work on many domains, such as online hate

mitigation targeting Basque and Spanish as well as crosslingual transfer with the aim of not only improving current techniques for CN generation, but also mitigating the lack of manually curated training data for many languages. For instance, while our experiments indicate that post-edited training data is required to attain reasonable performance, it is unclear how much additional such data would be a minimum for outperforming models trained on machine translated data. As manual annotation is costly and almost unattainable for gathering CN data, lessening the data requirement on expert effort is desirable. Furthermore, our work focuses on two languages as a first step towards examining data transfer and model transfer for CN generation. Future work is needed to investigate whether similar findings would hold in other languages or to pinpoint best practices for generating CN under low-resourced scenarios.

8. Acknowledgements

9. Ethical Statement

Tasks on Counter Narrative generation and corresponding datasets have been proposed by the scientific community to try to mitigate the spread of online hate. However, even good and legitimate purposes may imply certain risks, especially for people involved in the development of datasets

which may include extreme forms of Hate Speech. This includes researchers, annotators, data curators and evaluators involved in studies such as the one we present in this paper. Conscious of these facts, we took every precaution we could think of to avoid harmful effects.

10.   Bibliographical References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.

Victor de Boer, Michiel Hildebrand, Lora Aroyo, Pieter De Leenheer, Chris Dijkshoorn, Binyam Tesfa, and Guus Schreiber. 2012. Nichesourcing: harnessing the power of crowds of experts. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 16–20.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *CLiC-it*.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the international AAAI conference on web and social media*, 11(1):512–515.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Hossam Faris, Ibrahim Aljarah, Maria Habib, and Pedro A Castillo. 2020. Hate speech detection using word embedding and deep learning in the arabic language context. In *ICPRAM*, pages 453–460.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. T-projection: High quality annotation projection for sequence labeling tasks. *CoRR*, abs/2212.10548.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh

Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2018. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media*.

Pauline Möhle, Matthias Orlikowski, and Philipp Cimiano. 2023. Just collect, don't filter: Noisy labels do not improve counterspeech collection for languages without annotated resources. In *Proceedings of the 1st Workshop on Counter-Speech for Online Abuse (CS4OA)*, pages 44–61, Prague, Czechia. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Raj Ratn Pranesh, Ambesh Shekhar, and Anish Kumar. 2021. Towards automatic online hate speech intervention generation using pretrained language model.

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 (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pretrained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

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

Maria Estrella Vallecillo-Rodríguez, Arturo Montejo-Raéz, and Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71:227–245.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.