# Take Shelter, Zanmi:
# Digitally Alerting Cyclone Victims in Their Languages

**Nathaniel R. Robinson**
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland, USA
nrobin38@jhu.edu

## Abstract

Natural disasters such as tropical cyclones cause annual devastation and take a heavy social cost, as disadvantaged communities are typically hit hardest. Among these communities are the speakers of minority and low-resource languages, who may not be sufficiently informed about incoming weather events to prepare. This work presents an analysis of the current state of machine translation for natural disasters in the languages of communities that are threatened by them. Results suggest that commercial systems are promising, and that in-genre fine-tuning data are beneficial.

## 1 Introduction and Related Work

Natural disasters are among the most outstanding humanitarian crises in the 21st century (Iserson, 2014). The propensity of Atlantic hurricanes has been increasing in recent years and is likely to continue, due to effects of climate change (Hosseini et al., 2018). Hurricanes and cyclones can be particularly destructive. According to the US Office for Coastal Management,[1] tropical cyclones in the USA have cost over $1.3 trillion in property damage and caused over 6.8k deaths since 1980. They also tend to disproportionately harm socioeconomically disadvantaged populations, including countries and communities in lower income brackets, the socially isolated, and the physically and mentally impaired (Krichene et al., 2023). In the summer of 2008 alone, hurricanes hit Haiti, the poorest country in the Americas, and cost the country nearly $1 billion, or roughly 15% of its GDP at the time (Republic of Haiti, 2008). Cyclones also cause tragic loss of life. 2017's Hurricane Maria caused ~3,000 deaths in Puerto Rico and the Lesser Antilles (Baldwin and Begnaud, 2018).

As a preventative measure for these types of tragedies, political leaders often issue evacuation orders in at-risk areas in anticipation of an incoming storm (Younes et al., 2021). Other alerts are also commonplace via both news networks, government notices, and social media (Zhang et al., 2019). In many places civilians are encouraged to prepare for an incoming hurricane by storing food, filling automobile gas tanks, and securing hurricane shutters on windows (Rose, 2006). These activities take time and planning, and those who receive notice of these recommendations late may be underprepared.

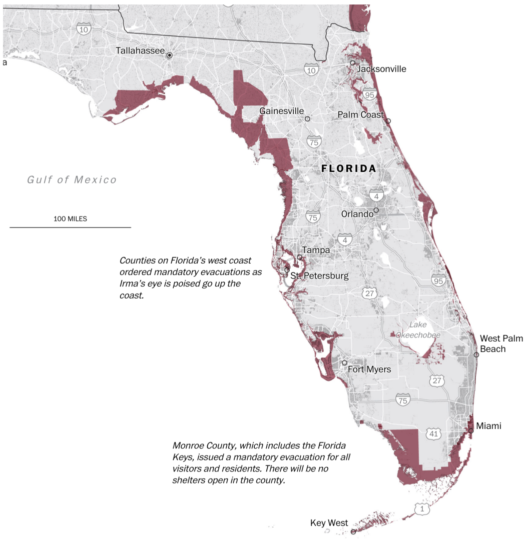
Tallahassee
Jacksonville
10
95
Gainesville
Palm Coast
75
FLORIDA
Gulf of Mexico
100 MILES
4
Orlando
Tampa
St. Petersburg
27
Counties on Florida's west coast ordered mandatory evacuations as Irma's eye is poised go up the coast.
Lake Okeechobee
West Palm Beach
Fort Myers
41
Miami
Monroe County, which includes the Florida Keys, issued a mandatory evacuation for all visitors and residents. There will be no shelters open in the county.
1
Key West

Figure 1: Approximate Hurricane Irma Mandatory Evacuation areas (red) as of 3:30pm on 10 September 2017. Image source: The Washington Post (Berkowitz et al., 2017). Evacuation areas in Lee and Collier counties were later expanded (Wong et al., 2018).

The effects of this phenomenon can vary by tropical storm. For instance, when Hurricane Irma struck Florida on 10 September 2017, the state's then-governor, Rick Scott, had already declared a state of emergency six days prior (Neuman, 2017). The six days' notice provided many civilians ample time to prepare for the storm. However, this also meant that those who prepared first claimed a majority of supplies early on. In the final days leading

[1] https://coast.noaa.gov/states/fast-facts/hurricane-costs.html

*Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 70–76
July 31, 2025

up to the storm, many grocery store shelves and gas station tanks were empty; leaving late preparers with few options (Reynolds and Collins, 2017). Florida's areas of mandatory evacuation due to Irma are illustrated in Figure 1. During the crisis of Hurricane Matthew, which struck Florida a year earlier, in 2016, then-US president Barack Obama had declared a state of emergency in the state only three days before the storm struck (Sarkissian et al., 2016). This gave civilians less time to prepare, but also resulted in a shorter period of resource scarceness. These decisions depend on numerous factors, including storm trajectory and speed (Regnier, 2020).

Since the majority of Floridan news reports, alerts, and social media posts are in English, English speakers were often the first to be informed of these crises; offering them a clear advantage over their non-English-speaking (predominantly Hispanic and Haitian) neighbors in claiming emergency preparedness resources (Tang, 2017).

Phenomena like these are part of why non-English-speaking communities in the USA are often the most harmed by natural disasters (Tang, 2017). Translation is needed to mitigate this, which can be expedited computationally when human translation resources are scarce or slow, via machine translation (MT). In this work, we explore two principal questions: (1) What MT systems are best applied in disaster scenarios, and under what circumstances? and (2) What MT model training practices contribute to success in this domain?

Other researchers have explored similar topics. In the wake of Haiti's devastating 2010 earthquake, there arose a renewed interest in Haitian MT for natural disaster relief and humanitarian aid (Margesson and Taft-Morales, 2010; Neubig and Hu, 2018). This interest inspired a task at the Workshop on Statistical Machine Translation (WMT) the following year: MT for the Haitian language (Callison-Burch et al., 2011). This task and the data set released with it led to subsequent works in Haitian MT (Stymne, 2012; Sennrich et al., 2016; Dholakia and Sarkar, 2014). Additional research has focused on MT for natural disaster communication (Cadwell et al., 2019), including multilingual systems that extend processing beyond translation (Sarioglu Kayi et al., 2020). We add to these previous works with a more current study focusing on MT into and out of English for four low-resource languages of cyclone-affected communities (Haitian, Jamaican Patois, Antillean Creole, and Mauritian Creole). We contribute:

- Indication that Google's commercial MT performs reasonably well on disaster text in our languages of focus
- Evidence that fine-tuning multilingual models on genre-appropriate data can improve natural disaster translation quality
- Evidence that generic mixed genre, or even religious discourse data is typically more helpful for training disaster-ready MT systems than Biblical data

## 2 Methodology

To analyze the state of MT of natural disaster alerts into the low-resource languages of affected areas, we evaluate on three test suites: **(1)** a set of Haitian SMS text messages sent during Haiti's 2010 earthquake with paired English translations; **(2)** scarcely available test sets for the languages of three other island nations affected by tropical cyclones; and **(3)** English corpora of tweets posted during cyclones, evaluated with back-translation pseudo-evaluation and human evaluation. We also conduct an exploration regarding what language and genre data is helpful for this MT application.

### 2.1 Haiti Earthquake SMS MT

First, to evaluate models' ability to translate disaster-related posts between English and Haitian, we employ the evaluation set from the Haitian MT task of 2011 WMT (WMT11) (Callison-Burch et al., 2011). This is a collection of 1.2k SMS messages that were sent during Haiti's 2010 earthquake, with both Haitian and English translations. In the set, as well as in the training and tuning sets corresponding to the same task, some personal identifiers are obfuscated by masks.[2]

In this evaluation, we compare three models that support Haitian translation: NLLB-200 (NLLB Team et al., 2022), an open-source multilingual model that supports a diversity of 203 languages; Kreyòl-MT (Robinson et al., 2024), a model specifically for Creole languages of the African diaspora that was fine-tuned with the Kreyòl-MT dataset from an mBART (Tang et al., 2021) initialization, and supports 41 languages; and Google Translate, a commercial MT system that supports 243

[2] First, last, and middle names; titles; phone numbers; and email addresses. These obfuscations were done by the original data authors, Callison-Burch et al. (2011).

| ***Experiment*** | *Covers multiple relevant languages* | *Evaluation on genre-appropriate data* | *Reference-based MT evaluation* |
|---|---|---|---|
| **Haiti Earthquake SMS MT** (§2.1) | ✗ | ✓ | ✓ |
| **Generic MT for Three Island Languages** (§2.2) | ✓ | ✗ | ✓ |
| **Pseudo-evaluation of Tropical Storm Alert MT** (§2.3) | ✓ | ✓ | ✗ |

Table 1: Summary of the features and limitations of the experimental methods outlined in sections 2.1, 2.2, and 2.3.

languages (as of August 2024).[3] We also experimented with post-trained versions of the open source models, as detailed in § 3.

Note that this first evaluation of **Haiti Earthquake SMS MT** can only give a far-from-comprehensive picture of low-resource disaster MT. Notably, it only deals with one language, though multiple low-resource language communities are frequently affected by tropical storms. Hence, we conduct further evaluations with more languages.

### 2.2 Generic MT for Three Island Languages

We evaluated translation into and out of English for three other languages of island nations frequented by cyclones: Jamaican Patois, French Antillean Creole (specifically Guadeloupean), and Mauritian Creole. Because these languages are not supported by NLLB-200, we used Google Translate and Kreyòl-MT only. And because French Antillean Creole is not supported by Google, we used the Haitian translation setting (the most closely related supported language) as an approximation. Due to data scarcity, translated evaluation sets in the news and social media genres for these languages are either nonexistent or too small to yield statistically significant results. Hence, we used the eval sets from the Kreyòl-MT dataset, which contain multiple genres.

### 2.3 Pseudo-evaluation of Tropical Storm Alert MT

The second evaluation of **Generic MT for Three Island Languages** gives language coverage, but lacks genre-appropriate eval data. So in our third and final evaluation, **Pseudo-evaluation of Tropical Storm Alert MT**, we used the same systems (Kreyòl-MT and Google Translate) to translate tweets from tropical storms. We used English corpora of such tweets from CrisisNLP's HumAID Dataset (Alam et al., 2021). Without ground truth translations, we were not able to compute MT scores directly. Instead we applied back-translation, translating from English into the target language and back into English, and then computing MT metric distances between the beginning and final English texts (to offer a rough approximation for MT quality). We employed the corpus' designated test set for a tropical storm that affected the area of each language: Hurricane Irma for Jamaican Patois, Hurricane Matthew for Haitian,[4] and Cyclone Idai for Mauritian Creole.[5]

Table 1 summarizes the different roles these three evaluations play, by displaying their contrastive features and limitations.

### 2.4 Fine-tuning experiments

Given the scarcity of genre-appropriate data sets for natural disaster applications in low-resource languages, we conducted an additional experiment to explore whether other genres of training text could still be helpful for this use case. We went about this by fine-tuning mBART for a single epoch (Tang et al., 2021) using different subsets of the Haitian training data from Kreyòl-MT. The four subsets were: (1) all 68,555 aligned sentences labeled as "Bible" genre; (2) exactly 68,555 aligned sentences labeled as "Religious" genre; (3) exactly 68,555 aligned sentences labeled as "Other/Mix" genre; and (4) all 1,072 aligned sentences labeled

[3] We demurred from including large generative language model systems such as OpenAI's ChatGPT, per the advice of Zhu et al. (2024); Robinson et al. (2023) that they are typically suboptimal for low-resource languages. This may be an interesting avenue for future work, as such models are frequently updated.

[4] We went back to evaluating Haitian instead of French Antillean Creole because of our back-translation pseudo-evaluation method, which would have yielded meaningless results if we used Haitian as a proxy with Google.

[5] Idai did not strike Mauritius, but it was the only Indian Ocean storm represented in the data.

| | hat→eng | | eng→hat | |
|---|---|---|---|---|
| | *chrF++* | *BLEU* | *chrF++* | *BLEU* |
| Kreyòl-MT | 39.9 | 22.1 | 37.2 | 19.0 |
| NLLB | 41.6 | 25.2 | 40.3 | 20.0 |
| Kreyòl-MT FT | 47.7 | 32.8 | 43.1 | 25.5 |
| NLLB FT | 45.8 | 30.0 | 43.7 | 24.0 |
| Google | **49.1** | **34.1** | **48.6** | **28.1** |

Table 2: Automatic scores for **Haiti Earthquake SMS MT**. Best scores **bold**.

| | jam–eng | | | | gcf–eng | | | | mfe–eng | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *chrF* | | *BLEU* | | *chrF* | | *BLEU* | | *chrF* | | *BLEU* | |
| | (→) | (←) | (→) | (←) | (→) | (←) | (→) | (←) | (→) | (←) | (→) | (←) |
| Kreyòl-MT | **77.0** | **94.6** | **66.9** | **92.3** | **69.5** | **71.1** | **53.6** | **54.3** | **65.4** | **57.7** | **52.7** | **40.0** |
| Google | 51.9 | 44.7 | 29.3 | 30.2 | 41.0 | 27.7 | 24.7 | 2.5 | 58.1 | 47.7 | 38.7 | 24.0 |

Table 3: Automatic scores for **Generic MT for Three Island Languages**. chrF is used in place of chrF++ to directly compare with results published by Robinson et al. (2024). Arrows →← indicate direction of translation.

as any of the genres "Narrative," "Wiki," and "Educational" added to 67,483 from the "Other/Mix" set (for equal train set sizes). The "Religious" genre differs stylistically from the "Bible" genre: while the latter is text directly from Bible translations, the former consists of recent religious discourses and publications. These genres were selected simply for data availability reasons. We evaluated all these fine-tuned models on the WMT11 test set.

## 3 Experimental Results and Conclusion

Table 2 displays automatic MT metrics chrF++ (Popović, 2017) and BLEU (Papineni et al., 2002) for **Haiti Earthquake SMS MT**, or translating the WMT11 test set between Haitian (hat) and English (eng). Fine-tuned (FT) models here were post-trained on the 16.7k aligned sentences in the SMS portion of the WMT11 train set (with the dev set for tuning). Kreyòl-MT was fine-tuned with an early stopping patience of 2 epochs, completing 5 epochs total. NLLB was fine-tuned twice (for each translation direction separately because it uses language-specific tokenizers) for 17 epochs, after which we observed dev accuracy did not improve.

In this evaluation, the closed-source commercial model performs best, followed by the fine-tuned Kreyòl-MT model. This highlights commercial models as strong choices for deployment of Haitian disaster MT, and suggests that the Creole-language-specific Kreyòl-MT model can surpass NLLB, when both are fine-tuned on in-genre data.

Table 3 shows our evaluation of **Generic MT for Three Island Languages**: i.e. using Kreyòl-MT eval sets for Jamaican Patois (jam), Guadeloupean Creole (gcf), and Mauritian Creole (mfe). Note that the Kreyòl-MT model has an advantage in this evaluation, since it was trained on data from the same sources as this test set.[6]

Table 4 shows our **Pseudo-evaluation of Tropical Storm Alert MT**: our back-translation-based eval from translating tropical cyclone tweet data from English into each of these island languages and then back for metric calculation. Google's system scores remarkably well in this scenario, though its particularly high scores in Jamaican-English translation are likely due to high similarity and duplication between source and translation text. (The Jamaican translation and English back-translation themselves have a high BLEU score of 49.4, indicating that the Jamaican translations are near copies of the source text, which inflates reconstruction score.) Kreyòl-MT by contrast performs relatively poorly on these automatic metrics.

A brief human evaluation for Haitian confirms this trend. We had a proficient Haitian speaker[7] blindly rate translations' accuracy and fluency for 100 utterances. From this evaluation, Google's average score was 4.55, where 5 signifies "no semantic errors, or like native"; and 4 signifies "semantic errors require minor fixes, or understandable

[6] This type of scenario is a frequent confounder in studies involving low-resource languages, and it highlights the need for more eval sets serving these language communities.

[7] ACTFL profiency *Superior*, as of 2019

|  | eng-hat |  | eng-jam |  | eng-mfe |  |
|---|---|---|---|---|---|---|
|  | *chrF++* | *BLEU* | *chrF++* | *BLEU* | *chrF++* | *BLEU* |
| Kreyòl-MT | 65.7 | 46.6 | 42.4 | 22.6 | 42.9 | 20.4 |
| Google | 79.2 | 65.8 | 95.3 | 90.7 | 83.0 | 69.3 |

Table 4: Back-translation reconstruction scores for **Pseudo-evaluation of Tropical Storm Alert MT**.

|  | hat→eng |  | eng→hat |  |
|---|---|---|---|---|
| mBART FT on... | *chrF++* | *BLEU* | *chrF++* | *BLEU* |
| 68.6k hat Bible | 8.4 | 0.6 | 18.2 | 4.0 |
| 68.6k hat Religious | 13.8 | 4.9 | 32.8 | 16.4 |
| 68.6k hat Other | 13.9 | 6.2 | 27.9 | 13.9 |
| 67.5k Other + 1.1k Narr./Ed./Wiki. | 13.6 | 6.3 | 15.0 | 6.5 |

Table 5: Additional study exploring fine-tuning corpus genre for WMT11 task

but not native." Kreyòl-MT's average score was 3.85, where 3 signifies "half or part of semantic information preserved, or disfluencies inhibit understanding." (For details see §A.) Further inspection after blind review revealed that Kreyòl-MT struggled particularly with named entities and Twitter characters such as '#' and '@'. This is understandable, since the Kreyòl-MT dataset does not contain much data from social media sources (Robinson et al., 2024), while Google's model may have been exposed to a large amount.

Table 5 displays the results of our additional experiment exploring post-training corpus genre (§2.4). The post-training corpora that achieved best performance here were the Other/Mix set and the religious set, suggesting that Biblical corpora may be less useful for MT in crisis scenarios.

In conclusion, humanitarian good can be accomplished by turning digital applications to help language communities in crisis. In our evaluation of MT in natural disasters for language communities that often fall prey to cyclones, we have found that commercial systems show promise for this application, that fine-tuning open source models on in-domain data improves results, and that mixed or discourse data is more beneficial for fine-tuning towards this task than Bible data.

## Limitations

One of this work's primary limitations is the lack of genre-appropriate evaluation sets for the languages included. This speaks to common difficulties in low-resource language technologies in general and the need for more resources. We also wish to remark that this work is primarily one of analysis and evaluation, intended to shed further light on the current state of MT for natural disasters in the languages of the areas that face them. Thus narrowly defined, the purpose of this paper is to point to future solutions to current problems that may be explored in greater depth.

## Ethics Statement

We wish to acknowledge briefly that the languages involved in this study are Creole languages with a colonial history. Creole languages are among the most marginalized and stigmatized both in technology, linguistics, academia, and society. Their speakers have historically been victims of colonial exploitation. The implications of this are twofold. First, Creole languages demand special attention, and we as a research community ought to take particular care to focus on Creole language needs, rather than neglect them. Second, and somewhat conversely, any research conducted with Creole languages ought to be approached with sensitivity and caution to avoid further exploitation or harm. It is our intention that this work might be of use to these communities and others burdened by natural disaster damage. It is not our intention to harm whatsoever, and if any content of this report happens to do so, we hope to be proactive in mitigating it to the extent possible.

## References

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021. Humaid: Human-annotated disaster incidents data from twitter. In *15th International Conference on Web and Social Media (ICWSM)*.

Sarah Lynch Baldwin and David Begnaud. 2018. Hurricane maria caused an estimated 2,975 deaths in puerto rico, new study finds. *CBS News*. Accessed: 2024-09-01.

Bonnie Berkowitz, John Muyskens, Tim Meko, Armand Emamdjomeh, Denise Lu, Aaron Steckelberg, Chiqui Esteban, Gabriel Florit, Ted Mellnik, and Chris Alcantara. 2017. What irma's wind and water did to florida. *The Washington Post*. Accessed: 2024-09-01.

Patrick Cadwell, Sharon O'Brien, and Eric DeLuca. 2019. More than tweets: A critical reflection on developing and testing crisis machine translation technology. *Translation Spaces*, 8(2):300–333.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the sixth workshop on statistical machine translation*, pages 22–64.

Rohit Dholakia and Anoop Sarkar. 2014. Pivot-based triangulation for low-resource languages. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 315–328.

SR Hosseini, M Scaioni, M Marani, et al. 2018. On the influence of global warming on atlantic hurricane frequency. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(3):527–532.

K. V. Iserson. 2014. Tackling the global challenge: humanitarian catastrophes. *The western journal of emergency medicine*, 15(2):231–240.

Hazem Krichene, Thomas Vogt, Franziska Piontek, Tobias Geiger, Christof Schötz, and Christian Otto. 2023. The social costs of tropical cyclones. *Nature communications*, 14(1):7294.

Rhoda Margesson and Maureen Taft-Morales. 2010. Haiti earthquake: Crisis and response. Library of Congress Washington DC Congressional Research Service.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Scott Neuman. 2017. Puerto rico prepares for category 4 hurricane irma. *NPR*. Accessed: 2024-09-01.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Eva D Regnier. 2020. What is six hours worth? the impact of lead time on tropical-storm preparation decisions. *Decision Analysis*, 17(1):9–23.

Republic of Haiti. 2008. Rapport d'évaluations des besions après désastre cyclones fay, gustav, hanna, et ike. `https://www.preventionweb.net/media/75380/download`. Accessed: 2024-09-01.

Shelby Reynolds and Ashley Collins. 2017. Hurricane irma: Families race to grocery, hardware stores for emergency supplies. *Naples Daily News*. Accessed: 2024-09-01.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Veronica Rose. 2006. Shutter protection for buildings in the florida building code. `https://www.cga.ct.gov/2006/rpt/2006-r-0645.htm`. Accessed: 2024-09-01.

Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown. 2020. Detecting urgency status of crisis tweets: A transfer learning approach for low resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4693–4703, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arek Sarkissian, J.D. Gallop, and Doug Stanglin. 2016. Hurricane matthew: Florida governor says, 'evacuate, evacuate, evacuate'. *USA Today*. Accessed: 2024-09-01.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sara Stymne. 2012. Clustered word classes for preordering in statistical machine translation. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 28–34.

Joanne Tang. 2017. Planning for non-english speakers in disaster situations. *RE: Reflections and Explorations: Volume 3*, page 163.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Stephen Wong, Susan Shaheen, and Joan Walker. 2018. Understanding evacuee behavior: A case study of hurricane irma. Technical report, Institute of Transportation Studies, UC Berkeley.

Hannah Younes, Aref Darzi, and Lei Zhang. 2021. How effective are evacuation orders? an analysis of decision making among vulnerable populations in florida during hurricane irma. *Travel behaviour and society*, 25:144–152.

Cheng Zhang, Chao Fan, Wenlin Yao, Xia Hu, and Ali Mostafavi. 2019. Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management*, 49:190–207.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

| | |
|---|---|
| 5 | = no semantic errors, or like native |
| 4 | = semantic errors require minor fixes, or understandable but not native |
| 3 | = half or part of semantic information preserved, or disfluencies inhibit understanding |
| 2 | = a few shared semantic themes with source, or some fluent elements present |
| 1 | = not a translation of the source, or not fluent at all, or wrong language |

Table 6: Combined scale for adequacy and fluency of translations

## A Human Evaluation

For simplicity in our human evaluation, we combined fluency and adequacy judgments into a single five-point scale, detailed in Table 6. We allowed our annotator to select scores in increments of 0.5 (i.e. 1, 1.5, 2, 2.5, etc.).