ITALERT: Assessing the Quality of LLMs and NMT in Translating Italian Emergency Response Text

Maria Carmen Staiano[1], **Lifeng Han**[2,4], **Johanna Monti**[3], **Francesca Chiusaroli**[1]

[1]University of Macerata, [2]LIACS & LUMC, Leiden University
[3]University of Naples "L'Orientale", [4]The University of Manchester
m.staiano@unimc.it, l.han@lumc.nl, jmonti@unior.it, f.chiusaroli@unimc.it

Abstract

This paper presents the outcomes of an initial investigation into the performance of Large Language Models (LLMs) and Neural Machine Translation (NMT) systems in translating high-stakes messages. The research employed a novel bilingual corpus, **ITALERT** (Italian Emergency Response Text) and applied a human-centric post-editing based metric (HOPE) to assess translation quality systematically. The initial dataset contains eleven texts in Italian and their corresponding English translations, both extracted from the national communication campaign website of the Italian Civil Protection Department. The texts deal with **eight crisis scenarios**: *flooding*, *earthquake*, *forest fire*, *volcanic eruption*, *tsunami*, *industrial accident*, *nuclear risk*, and *dam failure*. The dataset has been carefully compiled to ensure usability and clarity for evaluating machine translation (MT) systems in crisis settings. Our findings show that current LLMs and NMT models, such as ChatGPT (OpenAI's GPT-4o model) and Google Translate, **face limitations** in translating emergency texts, particularly in maintaining the appropriate register, resolving context ambiguities, and managing domain-specific terminology. The ITALERT corpus and evaluations are hosted openly at https://github.com/mcstaiano/ITALERT.

1 Introduction

LLMs have shown remarkable advancements in generating fluent and coherent translations. They are trained on large-scale multilingual datasets and can improve translation quality, efficiency and domain adaptation. The interest in LLMs also stems from the fact that they can provide valid translations for high-resource languages, producing competitive results with respect to traditional MT systems (Jiao et al., 2023).

Given their potential, this preliminary evaluation aims to assess the effectiveness of state-of-the-art language models during the preparedness and response phases of an emergency (Lindell et al., 2006) for the language pair **Italian→English** (**RQ**). The rationale for selecting this language pair lies in the need to communicate effectively with the non-Italian-speaking population, including immigrants, refugees, and tourists who are in Italy. The study focuses on identifying which translation system performs better under emergency conditions for specific natural disaster scenarios, providing critical insights to improve multilingual crisis communication in these phases.

The current Italian-English bilingual corpus consists of 13,218 words in total: 6,622 in Italian and 6,596 in English. It includes 440 segments across eight subdomains (flooding, earthquake, forest fire, volcanic eruption, tsunami, industrial accident, nuclear risk, and dam failure), extracted from the communication campaign website *Io non rischio*.[1] Table 1 provides an overview of the corpus, detailing the subdomains along with the word counts for both the original Italian texts and their English translations, sourced directly from the campaign's official website. In addition to the bilingual ITALERT corpus, we present our investigation methodology, and the evaluation of Google MT and ChatGPT using the HOPE metric (Gladkoff and Han, 2022).

2 Background and Related Work

Previous studies of crisis translation have demonstrated that MT systems can be fast, reliable tools for emergency response (Lewis, 2010). The past decade has seen a renewed focus on the possibility of using MT in preventing and mitigating disasters (Federici, 2016; O'Brien and Cadwell, 2017; Federici and Cadwell, 2018). However, few researchers in Italy have addressed the potential of

[1]https://www.iononrischio.gov.it

Proceedings of Machine Translation Summit XX Volume 1, pages 566–577, June 23-27, 2025

Subdomain	IT	EN
Flooding	633	637
Earthquake	368	372
Forest Fire	342	373
Volcanic Eruption	3231	3246
Tsunami	456	445
Industrial Accident	366	366
Nuclear Risk	735	682
Dam Failure	491	475
Total	**6,622**	**6,596**

Table 1: Subdomain-level word counts for the Italian-English bilingual corpus

MT systems in the crisis management workflow. During crises, it is crucial that messages are spread quickly and effectively to the population (Cadwell et al., 2019). But what impact can these messages have if they are communicated in a language that's foreign to the recipients or only partially understood by them?

Based on the latest tourism report provided by ISTAT(ISTAT, 2023), in 2023 Italy recorded 234.2 million overnight stays by foreign tourists, with non-residents making up 52.4% of total hospitality demand. At the same time, the country experienced a significant influx of immigrants and asylum seekers, representing more than 15 nationalities (ISTAT, 2024). In these multilingual and multicultural contexts, English is frequently used as a *lingua franca*, especially in interactions between migrants, institutions, and interpreters (Amato and Cirillo, 2024). This highlights the importance of using English as a vehicular language to foster mutual intelligibility in critical legal, medical, and social settings. For these reasons, we chose the language pair IT→EN for our translations. Our goal is to answer the following **research question**: How accurate are current language models in translating crisis-related texts from Italian to English?

In line with this direction, we started researching existing corpora on the topic of crisis translation in Italy, and we identified one dataset containing humanitarian response documents: HumSet (Fekih et al., 2022). However, this dataset was not selected for our study due to its lack of data in Italian, a critical requirement for our research objectives. While HumSet offers valuable multilingual resources (English, Spanish, French), the absence of Italian significantly limits its relevance to our analysis, which focuses on evaluating the quality of translations in crisis communication scenarios, specifically involving the IT→EN language pair.

Regarding MT evaluations and interpretability, previous work by Han et al. (2021) has examined both human and automatic evaluation methods. Recent research has also explored explainable MT evaluation (Leiter et al., 2024; Perrella et al., 2024), with a focus on providing detailed, interpretable error analyses. Additionally, the human-centric post-editing based metric HOPE offers both explainable feedback and supports the creation of a post-edited gold-standard corpus. As outlined by (Lommel et al., 2024), HOPE adopts a simplified and practical approach to human evaluation, specifically designed for machine translation outputs. Given that our study also aims to develop such a bilingual corpus, HOPE is a suitable choice, as it offers both detailed error analysis and the option to produce a post-edited corpus. The original HOPE metric has eight predefined error categories and severity levels. The eight error types in HOPE are: Impact (IMP), Required Adaptation Missing (RAM), Terminology (TRM), Ungrammatical (UGR), Mistranslation (MIS), Style (STL), Proofreading error (PRF), and Proper Name (PRN). The error severity levels and corresponding point values are: minor (1), medium (2), major (4), severe (8), and critical (16). For our annotation, we decided to use only 7 of the 8 error categories in HOPE, which will be explained in later sections (4.3).

3 Investigation Methodology

The ITALERT methodology for our investigation is shown in Figure 1.

1) The first step consists of data extraction and corpus collection. After the selection of the source texts in Italian, they were segmented into sentences for the MT evaluation phase. The source texts belong to eight subdomains (flooding, earthquake, forest fire, volcanic eruption, tsunami, industrial accident, nuclear risk, and dam failure) extracted from the national communication campaign website of the Italian Civil Protection Department (*Io non rischio*).

2) The second step involves selecting two MT systems and carrying out the automatic translation. Here, we aim to investigate Generative AI models (using ChatGPT-4o) and the standard NMT models (using Google MT) (Johnson et al., 2017). For ChatGPT, we used a zero-shot prompting technique

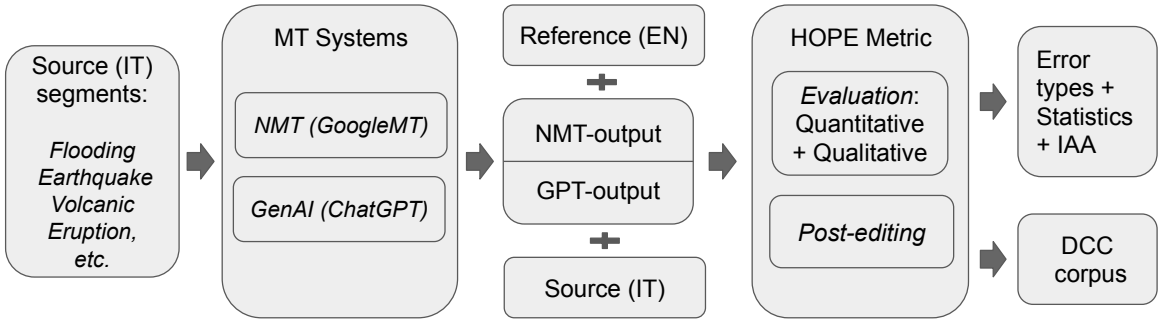

Figure 1: The ITALERT Investigation Methodology Design Framework.

(Cui et al., 2023), asking the model to perform a translation task. The *prompt* was as follows: "This is a document in Italian including crisis or emergency text for eight subdomains, specified in the heading part of each table. Can you please translate the content into English and keep the original document format in the tables?"

3) The third step is to carry out MT output assessment both quantitatively and qualitatively using the HOPE metric. For this step, we prepared two sets of triplets (source, reference, MT output) for both ChatGPT and Google MT (Johnson et al., 2017). Afterwards, post-editing was conducted on the reference gold standard in English, resulting in the creation of the **DCC** (Diamond Crisis Corpus) for ITALERT. [2]

3.1 Review of HOPE Metric calculations

HOPE uses error point scores to reflect how much effort is needed to post-edit a machine translation into a correct or gold-standard one. The error points are often annotated segment-by-segment (*e.g.,* a sentence as a segment). In its original formula, HOPE defines each segment as a translation unit (TU), and the Error Point Penalty of that unit (EPPTU) is calculated as the summation of each error type weighted by its severity level:

$$\text{EPPTU} = \sum_{i} Error_i \times Severity(i) \qquad (1)$$

where $Error_i$ is the type of error and $Severity(i)$ is the corresponding point value. Each translation unit is annotated independently from other units.

[2]We offered the reference set because of its availability. The HOPE metric itself does not require the "reference" set because it assumes the annotators understand the source and target language well, so the annotators can use the source to edit the system output, producing a post-edited gold standard.

At the system level, the HOPE metric can be calculated in two different ways. The first way is to sum all penalties for all j translation units, as below:

$$\sum_{TU_j} EPPTU_j = \sum_{i,j} Error_i \times Severity(i) \qquad (2)$$

This metric reflects the system- or document-level effort required to make the translated output correct. The second way is to calculate a similar metric using words as units instead of segments. This approach captures how many (or what percent) of words in the system output fall into the various error categories. In this study we chose the first option, using translation units to measure the system-level scores.

4 Experimental Work

4.1 Experimental Setup

For our experiment, we aimed to test ChatGPT as a representative of Generative AI and compare it with Google Translate as an NMT model, using the ITALERT dataset, as detailed in Table 1.

The evaluation of these two systems consisted of a two-step process. In the first step, we used the HOPE metric to evaluate the outputs. This metric returns both qualitative and quantitative analyses. Error annotation was performed by three professional linguists, native in Italian and proficient in English, all with a Master's Degree in Translation and a Bachelor's Degree in Linguistics. In the second step, post-editing was performed on the reference gold standard to create the DCC corpus. However, this version was not used during the error annotation process, as it will serve as a future resource for further testing and evaluation.

4.2 LLMs vs. NMT: comparative results on the crisis translation task

In this section we present: 1) descriptive statistics on error types and severities; and 2) a qualitative analysis of those errors.

Table 2 compares the performance of each system, listing the number of segments with error scores of 1, 2, 4, and >4, which we call minor, medium, major, and severe.

Error Type	Score	ChatGPT	Google MT
minor	1	104	109
medium	2	73	68
major	4	23	22
severe	>4	30	33
Total		**230**	**232**
Error rate		**0.52**	**0.53**

Table 2: Comparison of segment-level error counts for ChatGPT and Google MT on the crisis translation task. The error rate is computed over 440 total segments.

Examining this, we observe that:

- ChatGPT and Google MT present distinct patterns in segment-level error severity. Notably, ChatGPT produces more medium-severity errors (penalty score = 2) compared to minor ones (score = 1), whereas Google MT shows the opposite trend.
- Both systems exhibit a higher number of severe errors (score > 4) than major errors (score = 4), suggesting that when errors do occur, they often reach high levels of criticality. While the number of major errors is comparable (23 for ChatGPT and 22 for Google MT), Google MT presents slightly more severe errors (33 vs. 30), potentially raising concerns in high-impact applications.
- Overall, the total number of segments with error scores is 230 for ChatGPT and 232 for Google MT, resulting in error ratios of $230/440 = 0.52$ and $232/440 = 0.53$, respectively over 440 test segments. These values indicate that more than half of the evaluated segments contain non-trivial errors, underscoring the need for further system improvements to ensure reliability in sensitive domains such as crisis communication.

Table 3 reports the absolute error counts and percent error per category for ChatGPT and Google MT, based on a total of 611 and 728 annotated errors respectively.

- ChatGPT shows the highest number of errors in STL, TRM, and MIS, with fewer instances in IMP, PRF, UGR, and PRN.
- Google MT also shows the most errors in STL, MIS and IMP, followed by TRM and PRF. No errors were observed in the PRN category.
- The top error categories for both systems are STL, MIS, IMP, and TRM, indicating shared challenges across style, context, and terminology levels.
- The percent errors confirm STL as the most dominant category for both ChatGPT (40%) and Google MT (34%). For ChatGPT, TRM (25%) and PRF (8%) follow, whereas for Google MT, MIS (22%) and IMP (15%) are more prominent.

4.3 Inter-annotator agreement (IAA)

The annotation process was carried out by three professional linguists, native in Italian and proficient in English, all holding a Master's degree in Translation. After drafting a common set of guidelines, the annotators conducted an initial round of annotation to ensure consistency and a shared understanding of the annotation categories. During this phase, borderline cases and ambiguous instances were collected and discussed in a dedicated meeting. As a result of these discussions, the guidelines were refined and updated. In particular, the RAM (Required Adaptation Missing) category was merged into MIS (Mistranslation) to reduce overlap and improve clarity. Furthermore, a decision tree was developed to support the annotators in the classification process and facilitate decision-making for each of the categories.

The final IAA score was computed on a subset corresponding to 10 percent of the entire corpus. This evaluation aimed to assess the reliability and consistency of the annotation process after the consolidation of the guidelines. We measured IAA using well-established reliability measures commonly applied in computational linguistics research (Artstein and Poesio, 2008), drawing on prior work in MT evaluation and annotation reliability (Castilho, 2021). Our metrics include

Model	IMP	TRM	UGR	MIS	STL	PRF	PRN	Total Errors
ChatGPT	72 (11.7)	154 (**25.2**)	11 (1.8)	77 (12.6)	246 (**40.2**)	49 (**8**)	2 (**0.3**)	611
Google MT	111 (**15.2**)	134 (18.4)	19 (**2.6**)	162 (**22.2**)	248 (34)	54 (7.4)	0 (0)	728

Table 3: Absolute error scores for ChatGPT and Google MT in seven categories of errors from the HOPE model. The percent error is shown in parentheses. **Bold** indicates the highest percentage error in each column.

inter-rater agreement (IRR), Cohen's Kappa (Cohen, 1960) for pairwise comparisons, Fleiss' Kappa (Fleiss, 1971) and Krippendorff's Alpha (Krippendorff, 2011) for multi-annotator agreement.

Overall, these metrics confirm a high degree of annotation consistency across systems (Table 4). Both ChatGPT and Google MT reached strong levels of agreement according to multiple IAA metrics. Specifically, Percent Agreement was slightly higher for Google MT (IRR = 92.86) compared to ChatGPT (IRR = 90.48), suggesting that Google MT outputs might have been easier to classify in terms of error presence or absence. Similarly, Fleiss' Kappa and Krippendorff's Alpha confirmed substantial agreement levels for both systems, with Google MT again achieving marginally higher scores ($\kappa = 0.82$, $\alpha = 0.83$) than ChatGPT ($\kappa = 0.78$, $\alpha = 0.79$).

Table 5 presents the pairwise Cohen's Kappa scores for each combination of annotators. The results indicate substantial agreement across all pairs. Interestingly, annotators 1 and 2 showed the most agreement on Google MT ($\kappa = 0.916$), but performed significantly less well on ChatGPT ($\kappa = 0.076$), almost a 20-point difference. This suggests that some annotators encountered significant challenges across different outputs. Overall, IAA was lower on ChatGPT, whereas the higher pairwise agreement on Google MT, combined with its higher IRR scores, may indicate that its outputs were more predictable or less ambiguous in terms of error types *(e.g.* Style, Terminology, Proofreading, etc.), facilitating consistent judgments across annotators.

Model	IRR (%)	Fleiss κ	Kripp. α
ChatGPT	90.48	0.78	0.79
Google MT	92.86	0.82	0.83

Table 4: Inter-annotator coefficient scores for ChatGPT and Google MT: Inter-rater reliability, Fleiss' κ, and Krippendorff's α.

4.4 Evaluations: Levenshtein-perspective

To better understand the differences in string similarity between system outputs and reference translations at both character and word editing levels, we calculate the Levenshtein distance (Levenshtcin, 1966) for both systems against the reference translation on the English side.[3] The Levenshtein Distance (LevDis) measures the similarity or difference between two strings by counting the number of deletions, insertions, and substitutions required to transform one string into the other.

As in Table 6, the overall LevDis from ChatGPT is 11,812 compared to 10,544 from Google MT, across 440 segments. The average LevDis per segment is 29.82 for ChatGPT and 26.62 for Google MT, indicating that ChatGPT outputs are, on average, less similar to the reference strings than Google MT outputs. This is a very interesting outcome with two possible explanations. Assuming our human evaluation is correct and reliable, either LevDis is not a good metric to measure the text similarity, or the reference used has limitations.

Inappropriate metric: It is possible that LevDis is not an adequate or even appropriate metric to measure text similarity, especially semantic-wise, since it only matches the string similarity at the surface level. For instance, phenomena like negation can have a significant impact on language closeness, but are treated with equal weight as any other token in the LevDis calculation. Prior research supports this concern; for example, (Greenhill, 2011) argues that LevDis fails to identify language closeness in the tested data.

Limitations of the reference translation: It is also possible that relying on a single reference translation is limiting, as it may not adequately capture the variability and richness of natural language. Future work should therefore consider multi-reference evaluation settings to better account for this variation and provide a more robust assessment of translation quality.

[3] https://xlinux.nist.gov/dads/HTML/Levenshtein.html

Model	Annot. 1 vs 2	Annot. 2 vs 3	Annot. 1 vs 3
ChatGPT	0.76	0.76	0.84
Google MT	0.92	0.83	0.75

Table 5: Cohen's Kappa scores for ChatGPT and Google MT.

Levenshtein Dis	GPT-ref	Google-ref
Total Distance	11,812	10,544
Avg. dist./seg	29.82	26.62

Table 6: Levenshtein Distance scores comparing System Outputs (ChatGPT and Google MT) against the Original Reference

System	COMET	BLEU
Baseline: GPT	**88.83**	46.29
Google MT	88.73	**50.67***

Table 7: Evaluation results generated with MATEO. * indicates a significant difference with the first row (baseline).

4.5 Automatic Evaluation Metrics

To assess the overall performance of the systems under comparison, we employed two widely-used automatic metrics: BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) using MATEO (Vanroy et al., 2023). Due to missing references, 43 segments were excluded from the evaluation. All reported scores are based on the remaining 397 segments for which gold-standard reference translations were available. This filtering was applied consistently across systems to preserve comparability. All scores were computed with 1,000 bootstrap resampling iterations to estimate 95% confidence intervals, following SacreBLEU-compatible practices.

As shown in Table 7, Google MT significantly outperforms the GPT-based system in terms of BLEU, with a notable difference (50.67 vs. 46.29), indicating a higher degree of n-gram overlap with reference translations. On the other hand, COMET scores, which incorporate both reference and source-side information, are slightly higher for the GPT system (88.8 ± 0.7) compared to Google MT (88.7 ± 0.8). However, the difference is not statistically significant ($p = 0.2258$), indicating that both systems perform comparably on semantic evaluation. These results suggest that while GPT may retain a marginal edge in semantic coherence, Google MT is more aligned with reference outputs at the lexical level.

4.6 Qualitative Assessments of Two Systems

In this section, we categorize the errors made by ChatGPT and Google MT, providing examples based on the HOPE metric analysis presented in the previous section. Next, we summarise key aspects of the crisis domain MT, including the corpus itself and the performance of the translation systems. This is followed by a discussion to provide further insights into the challenges and considerations specific to this domain.

In Table 8, we report examples of stylistic errors found in ChatGPT's translation output. For instance, in the first sentence, ChatGPT translated "seminterrati" using "semi-basements". The term "semi-basements" is confusing and uncommon in English, which may compromise clear communication in a crisis context. In the second example, ChatGPT's use of "it may stop" introduces ambiguity, as "stop" could suggest a pause rather than a malfunction. In contrast, the reference term "get stuck" more effectively conveys the severity of the risk, especially in the context of elevator usage.

Further stylistic issues include the use of "safe spots" instead of "safe places," which feels less consistent with standard emergency communication terminology. Additionally, terms like "prohibit" and "hazards" may lead to confusion, as they are not part of everyday language and may be difficult for the general public to understand. These errors highlight how stylistic inaccuracies can impact the clarity and effectiveness of emergency communication, especially for lay readers.

In Table 9, we provide examples of mistranslations produced by Google MT. One critical error occurs in the translation of the sentence "utilizza un fermo per l'apertura degli sportelli dei mobili," which is rendered as "use a stopper to open the doors of the furniture", failing to convey the intended meaning. Additional errors, such as "closed space" and "order to stay away," are examples of linguistic calques from the Italian language. In

Source	ChatGPT	Google MT	Reference
Non scendere in cantine, seminterrati o garage per mettere al sicuro i beni: rischi la vita.	Do not go into basements, **semi-basements**, or garages to secure belongings: your life is at risk.	Do not go down into cellars, basements or garages to secure your belongings: you risk your life.	Do not enter cellars, basements, or garages to secure possessions; you will risk your life.
Evita l'ascensore: si può bloccare.	Avoid using the elevator: it may **stop**.	Avoid the elevator: it can get stuck.	Avoid the elevator: it may get stuck.
Individua i punti sicuri dell'abitazione dove ripararti in caso di terremoto.	Identify **safe spots** in your home where you can **take cover** during an earthquake.	Identify safe places in your home where you can take shelter in the event of an earthquake.	Identify safe places in your home to take shelter during an earthquake.
L'interdizione della zona rossa potrà durare molto tempo e l'accesso all'area sarà presidiato dalle forze dell'ordine.	Access to the red zone may be **prohibited** for a long time, and the area will be monitored by law enforcement.	The red zone ban may last a long time and access to the area will be supervised by law enforcement.	The red zone exclusion may last for a long time, and access to the area will be controlled by police forces.
Per questo, verifica che la tua casa sia sicura e adotta tutti gli accorgimenti necessari a eliminare le situazioni che possono rappresentare un pericolo.	Therefore, check if your home is safe and take all necessary precautions to eliminate potential **hazards**.	For this reason, make sure your home is safe and take all necessary precautions to eliminate situations that may represent a danger.	Therefore, verify that your home is safe and take all necessary measures to avoid situations that may pose a danger.

Table 8: ChatGPT Style Error Examples

Table 10, we present more error examples from different categories for both systems.

4.7 MT Challenges in Crisis Texts

Within the scope of our investigation, we believe that several challenges related to the crisis corpus and its translation represent valuable topics for discussion.

- *Literal vs context-appropriate translations:* In one example, the source text "Iniziano le operazioni di allontanamento delle persone con particolari necessità di assistenza sociosanitaria" was translated by Google MT as "Operations to **remove people** with particular needs for social and healthcare assistance begin". The phrase "to remove people" introduces a problematic lexical choice. In the context of a natural disaster or emergency, the expected term is "to evacuate," which carries a neutral connotation. By contrast, "to remove people" may imply coercion or force, potentially distorting the communicative intent of the source and undermining trust in the message. This example underscores the importance of context-aware lexical selection in high-stakes scenarios such as crisis communication. Another case is "Non scendere in cantine, seminterrati o garage durante l'alluvione," which was translated as "Do not go down into basements, **semi-basements**, or garages during the flood." The term "semi-basements" is rarely used in English and may confuse readers. These examples demonstrate how overly literal translations can reduce clarity and accessibility, particularly in crisis contexts. Adopting context-sensitive phrasing ensures better clarity and

Source	ChatGPT	Google MT	Reference
Da solo	On your own	**Alone**	On your own
In cucina, utilizza un fermo per l'apertura degli sportelli dei mobili dove sono contenuti piatti e bicchieri, in modo che non si aprano durante la scossa.	In the kitchen, use latches on cabinet doors containing plates and glasses to prevent them from opening during a tremor.	In the kitchen, **use a stopper to open the doors** of the furniture where plates and glasses are stored, so that they do not open during the earthquake.	In the kitchen, secure the cupboard flaps where plates and glasses are stored so they do not open during the earthquake.
Se sei in un luogo chiuso	If you are indoors	If you are in a **closed place**	Indoor
Iniziano le operazioni di allontanamento delle persone con particolari necessità di assistenza sociosanitaria.	The evacuation operations begin for people with specific social and healthcare needs.	Operations **to remove people** with particular needs for social and healthcare assistance begin.	In this phase evacuation operations begin for people with special social and health care needs.
Una volta diramato l'ordine di allontanamento, vai a casa e prepara la valigia.	Once the evacuation order is issued, go home and pack your suitcase.	Once the **order to stay away** has been issued, go home and pack your suitcase.	Once the evacuation order has been issued, go home, and pack your suitcase.

Table 9: Google MT Mistranslation Error Examples

accessibility in critical situations.

- *Impact of passive versus active voice on readers' understanding:* In Table 10, a Google MT Style Error is illustrated where one sentence is translated using the passive voice: "preventive actions that **can be taken**" instead of the more direct and active phrasing: "you can take preventive actions." We believe the choice of voice can significantly influence how recipients perceive and act upon the message. Active voice tends to be more transparent and more engaging, potentially making instructions easier to follow, especially in high-stakes scenarios typical of the crisis domain.

- *Terminology inconsistencies:* Errors in this category are particularly significant in the crisis domain, as they directly affect message accuracy and clarity. Notably, they emerge as the most frequent error type across both systems, as shown in the Table 3. Examples include "hazard" versus "danger," "closed space" versus "indoor," "voids" instead of "sinkholes," and "attention phase" instead of "alert phase". Addressing such discrepancies is essential to ensure precise and actionable communication during crises.

- *Meaning shift:* Some translation outputs from the tested systems result in meaning shifts or changes from the original text, which can have severe implications in the crisis domain. These examples include "false ceiling" *vs*. "suspended ceiling", "removal order" *vs*. "evacuation order".

- *Complex and long sentences:* The source corpus contains numerous long and complex sentences, which can hinder users' ability to process information effectively, leading to reduced actionability. We believe that instructions in crisis communication should be as concise and straightforward as possible for practical use. For instance, the source sentence: "In cucina, utilizza un fermo per

l'apertura degli sportelli dei mobili dove sono contenuti piatti e bicchieri, in modo che non si aprano durante la scossa." (as shown in Table 9), along with its reference, is lengthy and employs complex syntax. Simplifying and splitting such sentences would make them easier to process for both MT systems and end users, enhancing readability and usability in critical scenarios.

- *Register level*: Register plays a crucial role in ensuring that crisis communication is accessible and appropriate for its target users. We recommend using plain or lay language wherever possible to improve accessibility and comprehension. For instance, in the stylistic error highlighted in Table 10, both systems use the phrase "prohibit it," which is more formal, instead of the simpler and more commonly used lay term "ban it." Using plain language ensures that messages are accessible and easy to understand, particularly in high-pressure situations where clear communication is vital.

5 Conclusion and Future Work

In this study, we investigated the performance of LLMs and NMT systems in translating crisis-related texts. The evaluation was conducted using 440 segments from eight subdomains, with data sourced from the national communication campaign website *Io non Rischio*. ChatGPT-4o and Google Translate were selected as representatives of Generative AI and stand-alone NMT systems respectively, and were evaluated using a human-centric evaluation framework.

Errors from each system were categorised using the default 7 error types (merged from 8) from the HOPE metric, with a revised severity mapping, adjusted to account for the sensitivity of the crisis domain. The findings reveal that both systems share common error types but differ in their rankings. ChatGPT showed a high incidence of Style and Terminology errors, while Google MT was characterised by a greater presence of Mistranslation, Impact, Terminology, and Style issues. Importantly, both systems produced a non-negligible amount of severe and major errors, despite the predominance of minor and medium-level issues. The number of segments with severity ratings above 4 was slightly higher in Google MT outputs than in those of ChatGPT, indicating a greater incidence of critical errors. As seen in the qualitative analysis, several of these high-severity errors in Google MT translations had the potential to significantly distort the intended meaning and undermine the actionability of the messages. Interestingly, automatic evaluation metrics appear to diverge from human error analysis findings. While BLEU scores show a clear advantage for Google MT, indicating stronger surface-level fidelity, COMET scores are only marginally higher for the GPT system, suggesting comparable semantic adequacy. This trend aligns more closely with human judgments: HOPE-based error annotation reveals that Google MT's surface-level advantage does not correspond to improved quality in critical cases, as human annotators identified 728 errors in Google MT outputs, compared to 611 in those produced by ChatGPT. This discrepancy reflects limitations of using merely quantitative metrics in capturing context-sensitive, high-impact errors, and highlights the importance of complementary human-centric evaluations, especially in high-stakes scenarios (Hajek et al., 2024), as well as the need of developing domain specific automatic metrics in the future, e.g. for crisis translation.

In addition to the error analysis of current LLMs and NMT systems, using the HOPE framework, we also produced the Diamond Crisis Corpus (**DCC**), a new post-edited reference set derived from the ITALERT dataset.

Since ITALERT represents the first MT corpus on Italian crisis translation, we plan to add new subdomains, such as *public health emergencies* from the healthcare domain (Han et al., 2024). Other authoritative sources from which to extract the texts include Médecins Sans Frontières[4], The International Red Cross[5], and The United Nations Office for Disaster Risk Reduction[6]. We also plan to calculate the Levenshtein distance between the systems' outputs and the DCC corpus, to investigate differences in findings compared to those obtained using the original "gold" corpus. This analysis will help determine whether the post-edited corpus enhances the evaluation of translation performance.

Future work will explore in-domain training to address the challenges of context disambiguation and terminology management (Kirchhoff et al., 2011) and raise awareness of a responsible use of translation technology in high-stakes settings.

Finally, while we acknowledge that resources

[4] https://www.msf.org
[5] https://www.icrc.org/en
[6] https://www.undrr.org

comparable to ITALERT are not yet widely available, this work represents an initial step toward addressing the underexplored area of multilingual crisis communication (Cadwell et al., 2024), hoping it will serve as a foundation for future research and resource development in this emerging field.

Acknowledgments

This work was partially supported by the PhD programme in Humanities and Technologies funded by the University of Macerata under D.R. No 253/2023. We thank Professor Federico Federici from UCL for his valuable suggestions on the selection of texts for our corpus. We thank Kung Yin Hong (Kenrick) for helping with the HOPE metric and Levenshtein scores in the earlier stages of this work. We thank Willemijn Klein Swormink for the valuable discussion on a more interactive, visualised, and insightful database to build for the ITALERT project. We thank Serge Gladkoff, the CEO of Logrus Global LLC, for valuable advice on the manuscript. We thank Argentina Anna Rescigno and Antonio Castaldo for their contributions to the annotation process and for the brainstorming sessions, which helped us refine the annotation guidelines for the HOPE metric.

Sustainability statement

In this study, we conducted an evaluation of current LLMs and NMT systems for translation quality assessment within the crisis domain, without performing any training or fine-tuning. The computational requirements were minimal, as the evaluation involved translating only 440 segments.

References

Amalia Amato and Letizia Cirillo. 2024. Mediating english as a lingua franca for minority and vulnerable groups-introduction. *MEDIAZIONI*, (41):1–7.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

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

Patrick Cadwell, Sharon O'Brien, Aline Larroyed, and Federico M Federici. 2024. A crisis translation maturity model for better multilingual crisis communication. *INContext: Studies in Translation and Interculturalism*, 4(2):136–165.

Sheila Castilho. 2021. Towards document-level human mt evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yang Cui, Lifeng Han, and Goran Nenadic. 2023. MedTem2.0: Prompt-based temporal classification of treatment events from discharge summaries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 160–183, Toronto, Canada. Association for Computational Linguistics.

Federico M. Federici. 2016. *Mediating Emergencies and Conflicts*. Palgrave Macmillan, Houndmills.

Federico M. Federici and Patrick Cadwell. 2018. Training citizen translators: Design and delivery of bespoke training on the fundamentals of translation for new zealand red cross. *Translation Spaces*, 7(1):23–43.

Selim Fekih, Benjamin Minixhofer, Ranjan Shrestha, Ximena Contla, Ewan Oglethorpe, Navid Rekabsaz, et al. 2022. Humset: Dataset of multilingual information extraction and classification for humanitarian crises response. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4379–4389.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Simon J Greenhill. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4):689–698.

John Hajek, Yu Hao, Ambrin Hasnain, Anila Hasnain, Ke Hu, Maria Karidakis, Rachel Macreadie, Anthony Pym, and Juerong Qiu. 2024. Understanding and improving machine translations for emergency communications.

Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, 6:1211564.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

ISTAT. 2023. Nuovo picco delle presenze turistiche. Accessed: 2024-11-27.

ISTAT. 2024. Intensi flussi di immigrazione straniera, in lieve ripresa mobilità interna ed espatri. Accessed: 2024-05-28.

Wenxiang Jiao, Wenhui Wang, Jitong Huang, Xu Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint*, arXiv:2301.08745(1):1–10.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Katrin Kirchhoff, Anne M Turner, Amittai Axelrod, and Francisco Saavedra. 2011. Application of statistical machine translation to public health information: a feasibility study. *Journal of the American Medical Informatics Association*, 18(4):473–478.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

VI Lcvenshtcin. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, volume 10.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.

William Lewis. 2010. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

Michael K. Lindell, Ronald W. Perry, Carla Prater, and William C. Nicholson. 2006. Fundamentals of emergency management. page 485. FEMA, Washington, DC, USA.

Arle Lommel, Serge Gladkoff, Alan K Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, et al. 2024. The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 75–94.

Sharon O'Brien and Patrick Cadwell. 2017. Translation facilitates comprehension of health-related crisis information: Kenya as an example. *JoSTrans: The Journal of Specialised Translation*, 28:23–51.

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.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland. European Association for Machine Translation.

A Error Examples

We list in Table 10 additional error examples from Google MT and ChatGPT.

Source	ChatGPT	Google MT	Reference
Valuta bene se mettere al sicuro l'automobile o altri beni: può essere pericoloso.	Carefully consider if securing your car or other belongings is necessary, as it can be dangerous.	Consider carefully whether to secure your car or other **property**: it can be dangerous.	Carefully evaluate whether to secure your car or other possessions; it can be dangerous.
In questa fase, quando viene diramata un'allerta, è possibile compiere alcune azioni preventive per ridurre il rischio.	During this phase, when an alert is issued, you can take some preventive actions to reduce risk.	At this stage, when an alert is issued, there are some preventive actions that **can be taken** to reduce the risk.	You can take preventive actions to reduce the risk when an alert is issued.
Fai un elenco delle cose essenziali da portare con te (documenti, medicine, occhiali da vista, ecc.).	Make a list of essential items to bring with you (documents, medications, **glasses**, etc.).	Make a list of essential things to bring with you (documents, medicines, **glasses**, etc.).	Make a list of essentials to bring (documents, medicines, eyeglasses, etc.).
Prima di bere l'acqua dal rubinetto assicurati che ordinanze o avvisi comunali non lo vietino; lava bene i cibi che sono venuti a contatto con le ceneri vulcaniche.	Before drinking tap water, make sure municipal ordinances or advisories do not **prohibit** it; thoroughly wash food that has come into contact with volcanic ash.	Before drinking tap water, make sure there are no ordinances or notices **prohibiting** it; wash any food that has come into contact with volcanic ash thoroughly.	Before drinking water from the tap, make sure that municipal ordinances or notices do not **ban** it; wash well any food that has come into contact with volcanic ashes.

Table 10: Additional Error Examples from ChatGPT and Google MT