Improving MT-enabled Triage Performance with Multiple MT Outputs

Marianna J. Martindale† and **Marine Carpuat**‡
†College of Information, ‡Department of Computer Science
University of Maryland, College Park, USA

Abstract

Recent advances in Machine Translation (MT) quality may motivate adoption in a variety of use cases, but the success of MT deployment depends not only on intrinsic model quality but on how well the model, as deployed, helps users meet the objectives of their use case. This work focuses on a specific triage use case, MT-enabled scanning in intelligence analysis. After describing the use case with its objectives and failure modes, we present a user study to establish a baseline performance level and measure the mitigating effects of a simple intervention, providing additional MT outputs. We find significant improvements in relevance judgment accuracy with outputs from two distinct neural MT models and significant improvements in relevant entity identification with the addition of a rule-based MT. Users also like seeing multiple MT outputs, making it an appealing way to improve MT-enabled scanning performance.

1 Introduction

Recent years have seen dramatic advances in Machine Translation (MT) quality (Kocmi et al., 2022, 2023, 2024), making MT adoption in a variety of use cases all the more appealing. But intrinsic model quality does not dictate success or failure in MT deployment. For any given use case, the critical question is not how well the model performs on benchmark evaluations, but how effectively the model, as deployed, will help users accomplish their objectives. That requires understanding the objectives of the use case as well as the strengths and weaknesses of MT.

In this work, we focus on a triage use case, MT-enabled scanning in intelligence analysis, and its objectives and failure modes (Section 2.1). We will then discuss how the strengths and weaknesses of available MT systems may affect user performance (Section 2.2) and interventions that might improve performance (Section 2.3). Finally, we will detail our user study (Section 3) and provide recommendations for this and similar use cases based on the results (Section 4).

2 Background

2.1 MT-Enabled Scanning Use Case

In this work we refer to the process in intelligence analysis of labeling documents as relevant (to be kept for further analysis) or NTR (Nothing To Report) as “scanning". Like many triage use cases, scanning involves volumes of text large enough that it is impractical to have people who know the language perform triage. Instead, users familiar with the domain who do not know the language use MT to identify documents believed to be relevant enough to send for human translation. Because the users don’t know the language, they are susceptible to misleading errors in the MT output, but the risk of incorrect information from the MT output ending up in intelligence reports is mitigated by human translation before further analysis. However, MT errors that mislead the user still incur costs from irrelevant documents, wasting human translator time, or bear a risk of missing relevant documents.

2.2 Reliability of MT

Although there are no prior studies on MT-enabled triage for intelligence analysis, prior work on the reliability of MT can help us understand how the strengths and weaknesses of MT may affect this use case. Older MT approaches, such as statistical and rule-based MT (RBMT), suffered from fluency issues that can lead users to distrust the output (Martindale and Carpuat, 2018). The improved fluency of generated output comes with an increased risk of output that is detached from the meaning of the input, often referred to as hallucinations. This trend was initially observed in the earliest Neural MT (NMT) models (Koehn and Knowles, 2017; Lee et al., 2018; Martindale et al., 2019; Raunak et al., 2021) but has remained an issue in more recent MT models (Xu et al., 2023; Guerreiro et al., 2023) and Large Language Models (LLMs) (Kalai and Vempala, 2024). Despite their fluency, hallucinations may not be believable in context (Martindale et al., 2021), but if believable in context, the user will be misled. Without intervention, the user must rely on surface features such as fluency, document context, and real-world context in deciding whether the MT output is an accurate representation of the meaning of the source text.

2.3 Possible Interventions

There are many possible interventions that could reduce how often users are misled during MT-enabled triage tasks. Our interventions should help users calibrate their judgments of MT output to decrease the believability of errors while increasing the believability of accurate translations. Explainability approaches such as confidence scores may help users calibrate trust in AI models (Zhang et al., 2020), but users may still be misled by low-confidence incorrect output (Suresh et al., 2020) and can have difficulty detecting critical errors (Mehandru et al., 2023). For MT in particular, sentence-level confidence scores tend not to be well-calibrated without explicitly adapting the training to encourage better calibration (Kumar and Sarawagi, 2019; Wang et al., 2020; Lu et al., 2022) and lack the specificity needed to help the user decide which parts of the translation to believe. Fine-grained MT quality estimation (QE) approaches like those in WMT shared tasks on word-level QE (Specia et al., 2021) and fine-grained error span detection (Blain et al., 2023) provide additional information for the user, but the best models do not perform well enough and require considerable resources, with the top submissions in WMT23 only achieving F1 scores below 0.3 for models with as many as 13B parameters or ensembles of up to 12 models (Blain et al., 2023). Rather than simply highlighting error spans in the output, Briakou et al. (2023) improve explainability using contrastive phrasal highlights to draw the reader's attention to meaning differences. The approach was tested with bilingual users in a human translation quality review scenario, but monolingual users could apply linguistic resources such as dictionaries to the highlighted source text phrases to verify the severity of divergence. This is a promising approach, but it is unclear whether the current models are performant enough for deployment without significant engineering effort.

The ideal intervention can immediately be deployed with MT models of any quality and will have the potential to continue to help users even as newer, better models are deployed. The best fine-grained error detection model at WMT23 relied on pseudo-reference translations generated by off-the-shelf MT systems (Rei et al., 2023). What if we simply provided the user with the alternate translation? This type of intervention is appealing because it requires no additional data or specialized skills and can be used for any language where more than one MT system is available. Prior work has shown that displaying two MT outputs improves confidence and performance in MT-mediated communication without increasing cognitive load (Xu et al., 2014; Gao et al., 2015). We hypothesize that MT-enabled triage use cases can derive similar benefits.

3 User Study Design

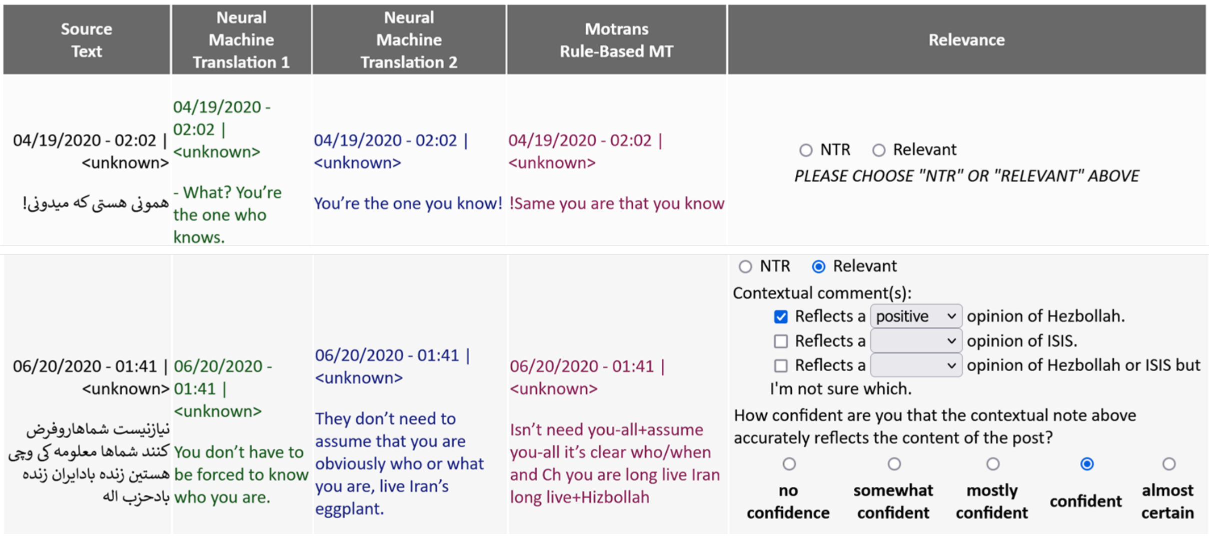
To establish a baseline risk level for MT-enabled triage in intelligence analysis and to measure the mitigating effects of practical interventions, we conducted a user study with Intelligence Analysts (IAs) from a US intelligence agency in the Washington, DC area with significant experience (at least three months) performing triage tasks with the aid of MT and little or no knowledge of the source language. In the next sections, we describe our interventions and design a scenario and tasks for the user study that mimic real triage tasks. We then address the format of the user study and analysis methods.

3.1 Intervention: Multiple MT Outputs

To mitigate the risks of misleading MT output, we propose two versions of the alternative translations intervention from Section 2.3 (pairing output from a single NMT system with output from a second NMT[1] system, and pairing a single NMT output with rule-based MT (RBMT) output). We also propose a combination of the two versions, displaying two NMT outputs with RBMT output.

IAs with output from only one MT system must rely on features of the output text, like fluency, and

[1] Note: LLMs were not yet available when the data for the user study tasks was translated and annotated. See Section 6



Source Text	Neural Machine Translation 1	Neural Machine Translation 2	Motrans Rule-Based MT	Relevance
04/19/2020 - 02:02 | <unknown> همونی هستی که میدونی!	04/19/2020 - 02:02 | <unknown> - What? You’re the one who knows.	04/19/2020 - 02:02 | <unknown> You’re the one you know!	04/19/2020 - 02:02 | <unknown> !Same you are that you know	○ NTR ○ Relevant *PLEASE CHOOSE "NTR" OR "RELEVANT" ABOVE*
06/20/2020 - 01:41 | <unknown> نیازنیست شماهاروفرض کنند شماها معلومه کی وچی هستین زنده بادایران زنده بادحزب اله	06/20/2020 - 01:41 | <unknown> You don’t have to be forced to know who you are.	06/20/2020 - 01:41 | <unknown> They don’t need to assume that you are obviously who or what you are, live Iran’s eggplant.	06/20/2020 - 01:41 | <unknown> Isn’t need you-all+assume you-all it’s clear who/when and Ch you are long live Iran long live+Hizbollah	○ NTR ● Relevant Contextual comment(s): ☑ Reflects a positive ⌄ opinion of Hezbollah. ☐ Reflects a ⌄ opinion of ISIS. ☐ Reflects a ⌄ opinion of Hezbollah or ISIS but I'm not sure which. How confident are you that the contextual note above accurately reflects the content of the post? ○ no confidence ○ somewhat confident ○ mostly confident ● confident ○ almost certain

Figure 1: The user interface for a Hezbollah/ISIS conversation thread.

contextual features, like plausibility, to decide the extent to which they believe an MT output reflects the meaning of the source text. A second MT output provides additional information to inform the decision. Differences between the two translations will draw attention to potential errors in fluent output and similarities between the translations can overcome disfluencies that would otherwise reduce the believability of an MT output.

In the second version of the intervention, RBMT output is not expected to provide the readability of neural MT output but does provide more interpretability than off-the-shelf NMT because every word or phrase in the output is a translation of specific words in the source. It is also easy to update with new named entities and specialized terminology, making it especially useful for keyword-spotting. For these reasons, the CyberTrans MT platform (Reeder, 2000) available to analysts throughout the US Intelligence Community includes Motrans RBMT for many languages (Martindale, 2012). Paired with one NMT output, Motrans can provide a similar effect to displaying a second NMT output if the output is sufficiently readable or contains relevant keywords. Paired with two NMT outputs with significant meaning differences, the Motrans output’s reliable connection to the source can make it a useful “tie-breaker".

3.2 User Study Tasks

This study focuses on Persian Farsi conversation threads in a scenario intended to be analogous to real intelligence analysis use cases. Persian was selected as the language for the study because it is of strategic importance and poses challenges for MT due to the limitations of available training data but there are open-source pre-trained models and commercial-off-the-shelf software available that can translate from Persian to English, as well as a Motrans capability. Conversation threads were chosen as our documents because, due to their difficulty, performance on conversation threads may be seen as a lower bound on analyst relevance judgment performance more broadly. Understanding any given message requires understanding its context in the conversation, and conversational text also often uses colloquial language which may be out of domain for MT systems.

For reasons of security and practicality, it is not possible to conduct the study using conversation threads from analysts’ actual data, so this study relies on an analogous collection of publicly available data gathered from user comments on Persian-language news articles. The topics for the user study are: Opinions related to the Russia-Ukraine conflict and Opinions related to terrorist organizations, specifically Hezbollah and ISIS. These topics were chosen because they relate to US intelligence priorities (strategic competition and violent extremist organizations) and are likely to elicit reactions among readers of Iranian news articles because of Iran’s support of Russia (Bowen et al., 2022) and Hezbollah (Humud, 2023) and Iran’s stance against ISIS (Arango and Erdbrink, 2014).

Analogous to the real MT-enabled triage use case, participants were asked to identify high-level features based on MT output in context. Each task consisted of one or more conversation threads that the user must scan for comments relevant to key intelligence questions, which they would label as either "Relevant" or "NTR" (Nothing to Report). They were also asked to identify information in the relevant comments as if they were adding a context note when passing the document to be translated. Finally, they were asked to rate their confidence in their judgments. A screenshot of the user interface for a Hezbollah/ISIS task conversation thread with both NMT outputs and Motrans RBMT output is shown in Figure 1, with the first comment unannotated and the second comment displaying the contextual note options.

The contextual note information was gathered in a multiple-choice, fill-in-the-blank style. Analysts could choose whether the comment is related to one or both of the relevant entities and whether the comment expresses a positive or negative opinion of that entity. Analysts could express uncertainty about the target of the comment by choosing an option that says they believe the comment reflects an opinion of one of the entities but they are not sure which. They could also express uncertainty about the stance of the comment by choosing "unclear" rather than "positive" or "negative." This allows for a granular evaluation of comprehension, from relevance judgment to information extraction to stance detection.

During the post-task survey, participants provided feedback validating the similarity of the tasks to their typical work, as discussed in Section 4.1.

3.3 Data and Annotation

The initial corpus of comment threads was collected in July 2022 by searching Persian-language news sites[2] for Farsi keywords related to the topics and then scraping the user comments, replies, and their publicly visible metadata (username, timestamp, and threading information) from the articles that were returned. Filtering for threads with at least two replies yielded 1,552 comments in 315 threads for Russia-Ukraine and 346 comments in 82 threads for the terrorism topic. Given limited annotation resources, we further filtered the Russia-Ukraine comments by selecting threads that were more likely to contain at least one comment with a potentially misleading translation in the context of this task using the Twitter-trained sentiment analysis model from TimeLMs (Loureiro et al., 2022) on the MTs of the comments and choosing threads that contained at least one comment for which the two NMT outputs had different sentiment labels. This resulted in 210 comments in 35 threads for annotation from the Russia-Ukraine topic.

The MT systems for the user study were chosen based on fitness for the use case. Because hallucinations are often tied to the training data (Raunak et al., 2021) we expect that output from a second model trained on different data is unlikely to produce the same hallucinations, so we want our NMT models to have been trained on substantially different data. One way to know the models were trained on different data is to use a bilingual model and a multilingual model, ensuring that even if both models were trained on similar Persian-English bitext, the multilingual model will have been exposed to additional English target text for other language pairs. To this end, we use a freely available massively multilingual pre-trained model, NLLB-200 (Koishekenov et al., 2022), and a commercial off-the-shelf system, SYSTRAN (version SPNS 9.7) as our two NMT systems. Open-source pre-trained models like NLLB-200 are appealing because they can be deployed on an intranet with minimal machine learning knowledge, and NLLB-200 is particularly desirable because it covers 200 languages, making it a logical choice for our baseline NMT system. SYSTRAN is a plausible second NMT system because it is familiar to US government users through long-standing collaboration with the Air Force (SYSTRAN, 2021) and previous integration in government translation platforms such as CyberTrans (Reeder, 2000).

The Persian Motrans capability that produced our RBMT outputs was developed from electronic dictionaries in the mid-2000s and continues to be updated with technical terms, named entities, and colloquialisms observed in sources such as news, technical documents, and web content. Motrans is optimized for adequacy rather than fluency. It handles ambiguity by providing alternative translations separated by a slash in the output and it attempts to split out of vocabulary tokens into smaller translatable words with '+' between the resulting translations in the output, as shown in the Motrans translation of the second comment in Figure 1. The ambiguous Farsi word کی is translated as *who/when*, and the incorrectly spaced phrase زنده بادحزب اله is translated as *long live+Hizbollah*.

Two Persian language analysts were recruited to provide gold standard annotations on the comments. The annotators completed the same relevance judg-

[2] isna.ir/news, tabnak.ir/fa/news, and khabaronline.ir/news

ment and contextual note task that user study participants would complete but using the source text rather than the MT output. They also evaluated the MT output quality using a task-focused adaptation of the evaluation scales from Licht et al. (2022). Each quality level was given a descriptive label to emphasize that they are labels rather than equally distanced points in a range. The lowest quality label was *MISS*, described as a translation that is so different from the meaning of the source text that a non-language-enabled analyst would not be able to reliably make even a relevance judgment. The second level was *REL-ONLY*, described as translation quality sufficient to make a relevance judgment but with significant information missing or incorrect. The third level was *GIST*, described as translating critical information correctly but with less important information missing or incorrect. Levels 4 and 5 were labeled *GOOD* and *EXCELLENT* respectively. Translations below GIST quality (MISS or REL-ONLY) can be considered potentially misleading in this scenario. Details can be found in Appendix A.

From the annotated comments, we selected conversation threads to use in two tasks per topic, each totaling approximately 15 comments. The threads were selected based on the relevance and MT quality judgments with the goal of including at least one unambiguously relevant comment per thread, at least two comments in each task where NMT1 was potentially misleading and NMT2 was GIST or better, and at least two comments where NMT1 was potentially misleading and Motrans was GIST or better. Of the 61 comments included in the user study, 32 were labeled relevant.

3.4 User Study Methods

The user study was conducted with a 2x2 design, with one between-subjects variable and one within-subjects variable. The between-subjects variable is whether the analyst sees one NMT output or two and the within-subjects variable is whether the analyst is provided rule-based MT output from Motrans in addition to the NMT output(s). Participating analysts were assigned to either the one-NMT or two-NMT condition and completed tasks both with and without Motrans output provided. The order of presentation of the conditions (with and without Motrans) was counterbalanced across analysts to control for ordering effects. Each analyst completed two tasks in each condition, and the order of all four tasks was counterbalanced to control for task-specific ordering effects.

The study was reviewed and approved by the University of Maryland Institutional Review Board, protocol number 1964637-1, and the Human Research Protection Program of the agency where the study took place. Informed consent was obtained from all participants prior to data collection.

Participants were recruited through messages on internal networking sites and mailing lists, with the goal of recruiting up to 40 qualified IAs. They were screened using a qualification survey, which also gathered relevant background information for qualified participants and asked about their perceptions of the MT they use. When they completed the background survey, their responses were validated against participation criteria, and if they qualified, they were asked to commit to completing the user study on a specific date and time of their choosing. Those who provided a date and time were assigned a batch of tasks round-robin style. In total, 35 IAs responded to the survey, and 26 completed the user study. Two of the survey respondents did not qualify because their MT use was in a language they knew and seven analysts did not respond to contact after the background survey. Two of the remaining 28 IAs, both from the 1-NMT condition, failed to complete the user study as assigned, leaving 12 participants in the 1-NMT condition and 14 in the 2-NMT condition.

4 Results

The discussion of the user study results is structured as follows. First, we validate the user study scenario and tasks. We then establish the baseline performance using output from one NMT system and demonstrate the mitigating effects on performance from providing additional outputs, followed by the effects on confidence. After summarizing these quantitative results, we briefly address acceptability of the interventions as indicated by responses on the pre- and post-task surveys.

4.1 Scenario Validation

Responses in the post-task survey verified whether the user study tasks were similar to intelligence analysis foreign language triage tasks. No participants said the tasks were “Much Easier" or “Much Harder" while 15% of users said that they were easier than their “typical foreign language text triage tasks," 42% said they were of similar difficulty, and 42% said they were harder. In open ended-

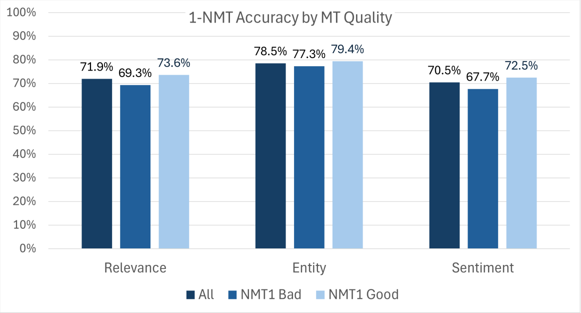


Figure 2: Mean accuracy in the 1-NMT condition across all examples (All) and with NMT1 quality below GIST (NMT1 Bad) and GIST or better (NMT1 Good), showing highest accuracy for entities and lowest for sentiment with a 4.8 point difference in sentiment accuracy between NMT1 Good and Bad.

responses regarding the elements of the user study that were similar to or different from analysts' typical triage tasks, the most frequently mentioned similarity was the overall framing, mentioned by 11 analysts. Five analysts mentioned similar MT quality and four analysts mentioned similar task difficulty. Three analysts noted that similar to their tasks, the conversations lacked context and the comments were short and informal. However, two analysts cited the length of the text as a difference, noting that they typically triage whole documents. Other differences that were mentioned included topic (seven analysts), language (three analysts), and their familiarity with the topic (five analysts). Only one analyst mentioned a difference in the structure of the task, stating that they do not typically write contextual notes "but its [sic] a good idea." Overall, these responses indicate that the scenario for the user study is comparable to many analyst workflows and the conversation threads selected are analogous to at least some real-world MT-enabled triage use cases.

4.2 Baseline Analyst Performance

Relying on output from only one NMT system, users (n=12) averaged 70% or higher accuracy on all three levels of comprehension, as seen in Figure 2. The entity accuracy score is highest (nearly 80%), likely because it is often possible to quickly tell when a comment refers to an entity by spotting the entity's name. The sentiment score is lowest (70.5%), supporting the intuition that identifying the stance towards the subject of a comment is more difficult than just identifying the subject of the comment.

Partitioning the comments based on the quality of the output from NMT1, we can measure performance on the potentially misleading examples (*NMT1 Bad*) compared to the *NMT1 Good* examples. For all three accuracy measures (relevance, entity, and sentiment), we see that mean accuracy is lower when the MT quality is bad (below GIST) and higher when the quality is good (GIST or better). The biggest difference is in sentiment, where the mean accuracy for bad translations is 67.7% compared to 72.5% for good translations.

The overall baseline accuracy reflects the utility of the baseline NMT system for this triage task, but leaves significant room for improvement, even when the MT output is fairly high.

4.3 Impact of Interventions on Accuracy

As described in Section 3.4, the user's response to each comment provides three labels we can score for accuracy: relevance judgment, sentiment for Entity A, and sentiment for Entity B. Each user's responses were compared against the gold standard annotations. We want to see whether adding a second NMT, adding RBMT, or the combination of both significantly affects relevance, entity, or sentiment accuracy, so we build three Generalized Linear Mixed Effects Models (GLMM) (one with each type of accuracy as the response variable) with fixed effects for the presence of a second NMT, presence of RBMT, and interaction between NMT and RBMT. We used random effects to control for user and item. For each GLMM, there are 1586 observations, grouped by item (61) and user (26).

Results of the GLMM with Relevance Accuracy as the response variable are shown in Table 1. We see a significant ($p < 0.05$) increase from adding a second NMT (OR=2.26, CI=1.35-3.85, p=0.0032) as well as adding RBMT (OR=1.52, CI=1.37-3.74, p=0.041) and a significant interaction from providing both (OR=0.52, CI=0.29-0.91, p=0.03). Based on this odds ratio, a hypothetical analyst with 3:1 odds of being correct in their relevance judgments with only one NMT would have their odds increased to 6.8:1 with a second NMT output. The same analyst would have their odds increased to 4.5:1 with the addition of RBMT. Note the negative β value for the interaction between NMT and RBMT. This means that although we would expect adding both a second NMT and RBMT to increase the analyst's odds of being correct to 10.2:1, the interaction effect means the odds only increase to 5.3:1, which is higher than just adding RBMT but

Coefficient	β	Odds Ratio	Confidence Interval	p
(Intercept)	1.497	4.469	2.476 - 8.067	< 0.001
2-NMT	0.816	2.262	1.370 - 3.735	0.0014
w/ RBMT	0.416	1.516	1.018 - 2.259	0.0408
2-NMT+RBMT	-0.660	0.517	0.294 - 0.909	0.0219

Table 1: GLMM for Relevance Accuracy showing largest significant (p<0.05) effect from the second NMT.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.479	1.614	0.719 - 3.627	0.3627
w/ RBMT	0.554	1.740	1.204 - 2.514	0.0032
2-NMT+RBMT	-0.505	0.603	0.362 - 1.007	0.0530

Table 2: CLMM for Entity Accuracy showing significant improvement (p<0.05) from adding RBMT.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.456	1.578	0.433 - 1.485	0.1540
w/ RBMT	0.251	1.285	0.933 - 1.770	0.1250
2-NMT+RBMT	-0.331	0.718	0.460 - 1.122	0.1460

Table 3: CLMM for Sentiment Accuracy showing no significant effects.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	1.052	2.863	1.098 - 7.466	0.0315
w/ RBMT	0.037	1.038	0.780 - 1.382	0.7980
2-NMT+RBMT	0.098	1.103	0.755 - 1.612	0.6130
Relevance Accuracy	0.349	1.417	0.924 - 2.174	0.1100
Entity Accuracy	-0.404	0.667	0.264 - 1.691	0.3940
Sentiment Accuracy	1.076	2.931	1.518 - 5.660	0.0014

Table 4: CLMM with Confidence showing significant (p<0.05) effects from Sentiment Accuracy.

lower than just adding the second NMT.

For entity accuracy (Table 2), we see a significant (p<0.05) improvement from adding RBMT ($exp(\beta)$=1.74, CI=0.186-0.922, p=0.0089). Based on this $exp(\beta)$, an analyst with 3:1 odds of being either iffy or right would increase their odds to about 5.2:1. We see a similar effect size for adding a second NMT, but it is not statistically significant, and the 95% confidence interval ranges from a detrimental 0.7 to a dramatic odds improvement of 3.6, so we cannot draw conclusions on the effect of a second NMT on entity accuracy. Once again, we see a negative β for the interaction between adding a second NMT and RBMT, although it is not significant.

For sentiment accuracy (Table 3), we see no statistically significant effects with adding a second NMT or RBMT (p>0.1). Sentiment is the deepest level of comprehension in this user study, so it was the least likely to be improved with the addition of a second NMT and/or RBMT. Sentiment judgment is beyond the scope of typical MT-enabled triage tasks, and these results show that adding a second NMT and/or RBMT does not improve accuracy reliably enough to suggest that the scope of MT-enabled triage should be expanded to include tasks at the level of sentiment judgment without oversight by analysts that know the language.

4.4 Effects of Interventions on Confidence

In addition to measuring accuracy, we also track self-declared user confidence. To assess the impact of adding RBMT and/or a second NMT on analyst confidence, we fit four additional models.

Coefficient	β	Odds Ratio	Confidence Interval	p
(Intercept)	1.665	5.287	2.930 - 9.545	3.2e-8
2-NMT	0.577	1.781	1.053 - 3.013	0.0315
w/ RBMT	0.445	1.560	1.038 - 2.345	0.0322
2-NMT+RBMT	-0.685	0.504	0.284 - 0.895	0.0194
Confidence	1.176	3.241	1.898 - 5.534	1.7e-5

Table 5: GLMM for Relevance Accuracy with Confidence, indicating well-calibrated Confidence.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.283	1.327	0.602 - 2.924	0.4823
w/ RBMT	0.549	1.732	1.194 - 2.512	0.0038
2-NMT+RBMT	1.181	3.258	1.194 - 5.409	4.9e-6
Confidence	-0.514	0.597	0.357 - 1.001	0.0503

Table 6: CLMM for Entity Accuracy with Confidence, showing significant effects (p<0.05) from RBMT and interaction with NMT and RBMT.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.246	1.278	0.707 - 2.311	0.4162
w/ RBMT	0.233	1.263	0.913 - 1.746	0.1584
2-NMT+RBMT	-0.353	0.702	0.448 - 1.101	0.1233
Confidence	1.321	3.748	2.420 - 5.800	3.2e-9

Table 7: CLMM for Sentiment Accuracy with Confidence, indicating well-calibrated confidence.

Following the pattern of the previous models, we fit a cumulative link mixed effects model (CLMM) with confidence as the response variable and second NMT, RBMT, and their interaction as fixed variables. We also added relevance accuracy, entity accuracy, and sentiment accuracy as fixed variables. This model shows whether each of these features (presence of each intervention and each type of accuracy) is a good predictor of the user's confidence.

As shown in Table 4, we observe a large increase in odds of higher user confidence from adding a second NMT output ($exp(\beta)$ =2.86, CI=1.098 - 7.466, p=0.032).Adding RBMT does not have a significant effect, and no significant interaction is observed. Relevance and Entity accuracy do not have a significant effect on user confidence, but Sentiment accuracy has a large statistically significant effect ($exp(\beta)$=2.93, CI=1.518 - 5.660, p=0.008), nearly tripling the odds of higher confidence with higher sentiment accuracy. This may indicate that sentiment judgment was front-of-mind when users chose their confidence level.

If analyst confidence is well-calibrated with analyst accuracy, it should be true that not only is accuracy a strong predictor of confidence but confidence is also a strong predictor of accuracy. Given that sentiment accuracy is a stronger predictor of analyst confidence than the presence of a second NMT and/or RBMT, we suspect that analyst confidence is reasonably well calibrated with at least sentiment accuracy. We can directly test this by adding confidence as another fixed effect in the relevance, entity, and sentiment accuracy models and comparing the results.

With confidence added to the relevance accuracy model as a fixed effect, we see minimal change in the effect of RBMT as shown in Table 5, but the odds ratio for adding a second NMT drops from 2.26 to only 1.78. Confidence is a strong predictor of relevance accuracy (OR=3.24, CI=1.898 - 5.534, p=4.95e-5), and the model with the confidence fixed effect is also a significantly better (p<0.01) model based on AIC (1310.8 vs 1335.8) and log-likelihood (-645.39 vs -661.92). The large confidence effect and model improvement suggest that analyst confidence is well-calibrated to relevance accuracy.

Adding confidence to the entity model (Table 6)

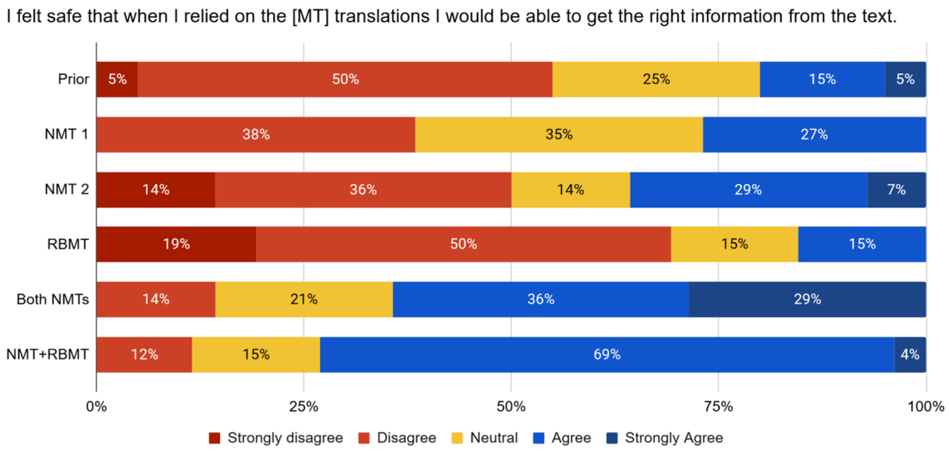


Figure 3: Participant responses to the safety item.

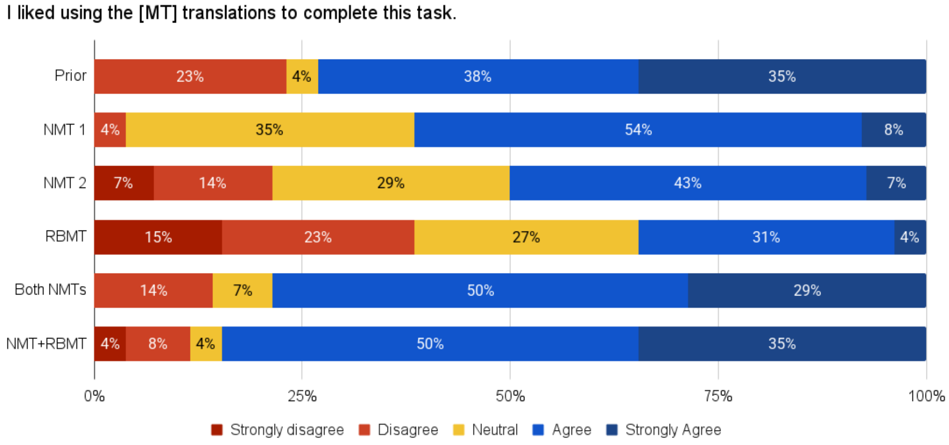


Figure 4: Participant responses to the likability item.

results in minimal change to the effects of adding a second NMT or RBMT, but confidence is as strong a predictor of entity accuracy as it was for Relevance accuracy (OR=3.26, CI=1.194 - 5.409, p=3.4e-5) and the entity accuracy model with confidence also demonstrates significant (p<0.01) improvements in AIC (1797.4 vs 1807.6) and log-likelihood (-882.68 vs -896.80) to those observed in the relevance model with the confidence fixed effect, suggesting that confidence is also well-calibrated with entity accuracy.

As with the original sentiment accuracy model, we see no significant effects from adding RBMT or a second NMT output (Table 7). Confidence is the only fixed effect to have a significant effect on sentiment accuracy (OR=3.75,CI=2.42-5.80, p=2.2e-8), verifying that just as sentiment accuracy is a strong predictor of confidence, confidence is also a strong predictor of sentiment accuracy. We also see significant (p<0.01) improvements to the AIC (2474.3 vs 2507.2) and log-likelihood (-1226.2 vs -1246.6) of the model from adding the confidence effect. This tells us that even when adding RBMT or a second NMT output does not affect accuracy, it also does not hurt confidence calibration.

4.5 User Feedback

For a mitigation to be effective, users must be willing to accept the resulting system. Key responses from the survey are the questions about likability and safety.

Figures 3 and 4 show how safe analysts felt when relying on the combinations of MT output and how much they liked using each combination. Less than 50% of participants agreed that they felt safe they would be able to get the right information from the text using the MT they typically have access to on-the-job or any one MT system from the user study.

With both NMTs, 65% of participants felt safe and 73% felt safe with NMT and RBMT. They liked having two NMT outputs (89%) but did not like using the NMT2 output as much as NMT1 (50% and 61%, respectively), and even though only 35% liked using RBMT, 85% liked using both NMT and RBMT. These seeming contradictions may be tied to how the analysts see themselves using the MT. Analysts may feel safe that they can get the right information because they believe they will be able to evaluate the information effectively. Similarly, analysts seem to like having access to RBMT as long as they have something to compare against. Prior work has indicated that IAs may be more likely than the general population to have an internal locus of control (Crouser et al., 2020), and that could explain their confidence that they will be able to take advantage of less-than-ideal MT output. Their open-ended responses give some insight as to how they use these combinations, with six analysts mentioning using Motrans for keyword spotting. As one analyst put it, “I used the literal translations very sparingly; mostly for the literal translation of a word, which I then plugged into the right spot of the neural translations."

5 Conclusions

We conducted a user study to establish a baseline level of IA performance on MT-enabled triage tasks and to measure the potential mitigating effects of a simple intervention, providing additional MT outputs. The user study found significant improvements in relevance judgment accuracy with output from two distinct NMT models and significant improvements in relevant entity identification with the addition of Motrans RBMT. The availability of additional MT outputs had little effect on analyst accuracy for the task that required the deepest comprehension of the text, identifying the sentiment towards the identified entity. Adding Motrans RBMT output had little effect on analyst confidence, but providing a second NMT output significantly improved it. This does not appear to be overconfidence, as confidence remained a strong predictor of accuracy across all three types of accuracy. Analysts also expressed a preference for seeing multiple MT outputs even when they felt that NMT1 provided better translations and praised the availability of multiple outputs in their open-ended post-task survey responses.

6 Recommendations and Future Work

Based on the analysts’ preferences and the improvements in relevance judgment accuracy, we recommend that two MT outputs be displayed side-by-side wherever IAs conduct MT-enabled triage. RBMT such as Motrans, which can be rapidly updated with new named entities and technical terms, can help analysts with keyword spotting when the NMT misses them, but a second NMT may provide more benefit to relevance judgment overall. If it is practical to provide outputs from two NMT systems that are sufficiently different in model architecture and/or training data, users can benefit from the readability of the NMT while also gaining the ability to triangulate meaning between the two outputs. Some MT systems (including SYSTRAN) provide the ability to integrate terminology lists, which could replicate the entity recognition benefits of the RBMT system with the fluency of NMT.

However, we caution that despite the significant improvements to relevance judgment accuracy from providing multiple MT outputs, this should not be taken as evidence that these interventions will allow analysts to perform tasks using MT output that require higher levels of comprehension than triage. The lack of significant improvement in sentiment accuracy supports maintaining the status quo of not reporting off MT output without verification by a language-enabled analyst.

This study began before LLMs were available, leaving several open opportunities for future work. Rather than using two NMT models, a single LLM could be used to produce more than one translation, as Gero et al. (2024) did with a variety of sensemaking tasks. LLMs can also be prompted to post-edit (e.g., Xu et al., 2024; Raunak et al., 2023; Chen et al., 2024; Vidal et al., 2022; Ki and Carpuat, 2024) or provide quality estimation (e.g., Huang et al., 2024; Rei et al., 2023; Fernandes et al., 2023). Further work is needed to determine the optimal way to use these approaches to benefit MT-enabled triage use cases.

Additionally, more user testing is needed to determine ways to effectively display multiple translations of longer text. The benefits of the second MT output may be outweighed by the difficulty in actually comparing those outputs if long translations are just dumped into adjacent text boxes.

Acknowledgments

We gratefully acknowledge the invaluable quantitative analysis feedback and guidance from Susannah B. F. Paletz.

References

Tim Arango and Thomas Erdbrink. 2014. U.S. and Iran Both Attack ISIS, but Try Not to Look Like Allies. *The New York Times*.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 Shared Task on Quality Estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Andrew S. Bowen, Clayton Thomas, and Carla E. Humud. 2022. Iran's Transfer of Weaponry to Russia for Use in Ukraine. CRS Report IN12042, Congressional Research Service.

Eleftheria Briakou, Navita Goyal, and Marine Carpuat. 2023. Explaining with Contrastive Phrasal Highlighting: A Case Study in Assisting Humans to Detect Translation Differences. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11220–11237, Singapore. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative Translation Refinement with Large Language Models. *arXiv preprint*. ArXiv:2306.03856 [cs].

R. Jordan Crouser, Alvitta Ottley, Kendra Swanson, and Ananda Montoly. 2020. Investigating the role of locus of control in moderating complex analytic workflows. *EuroVis 2020-Short Papers*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 852–863, Vancouver, BC, Canada. Association for Computing Machinery.

Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–21, New York, NY, USA. Association for Computing Machinery.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the Source Language: How Large Language Models Evaluate the Quality of Machine Translation. *arXiv preprint*. ArXiv:2401.06568 [cs].

Carla E. Humud. 2023. Lebanese Hezbollah. CRS Report IF10703, Congressional Research Service.

Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, Vancouver BC Canada. ACM.

Dayeon Ki and Marine Carpuat. 2024. Guiding Large Language Models to Post-Edit Machine Translation with Error Annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273, Mexico City, Mexico. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Workshop on Neural Machine Translation*, Vancouver, BC. ArXiv: 1706.03872.

Yeskendir Koishekenov, Vassilina Nikoulina, and Alexandre Berard. 2022. Memory-efficient NLLB-200: Language-specific Expert Pruning of a Massively Multilingual Machine Translation Model. *arXiv preprint arXiv:2212.09811*.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of Encoder Decoder Models for Neural Machine Translation. *arXiv preprint*. ArXiv:1903.00802 [cs, stat].

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in Neural Machine Translation.

Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent Human Evaluation of Machine Translation across Language Pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA. Association for Machine Translation in the Americas.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning Confidence for Transformer-based Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2353–2364, Dublin, Ireland. Association for Computational Linguistics.

Marianna Martindale and Marine Carpuat. 2018. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, MA. Association for Machine Translation in the Americas.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 233–243, Dublin, Ireland. European Association for Machine Translation.

Marianna Martindale, Kevin Duh, and Marine Carpuat. 2021. Machine Translation Believability. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 88–95.

Marianna J. Martindale. 2012. Can Statistical Post-Editing with a Small Parallel Corpus Save a Weak MT Engine? In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2138–2142, Istanbul, Turkey. European Language Resources Association (ELRA).

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for Automatic Translation Post-Editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Florence M. Reeder. 2000. At Your Service: Embedded MT As a Service. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 Shared Task on Quality Estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 689–730, Online. Association for Computational Linguistics.

Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, pages 315–324, New York, NY, USA. Association for Computing Machinery.

SYSTRAN. 2021. Government Translation Solutions | SYSTRAN Technologies.

Blanca Vidal, Albert Llorens, and Juan Alonso. 2022. Automatic Post-Editing of MT Output Using Large Language Models. pages 84–106.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the Inference Calibration of Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Bin Xu, Ge Gao, Susan R. Fussell, and Dan Cosley. 2014. Improving machine translation by showing two outputs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3743–3746.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

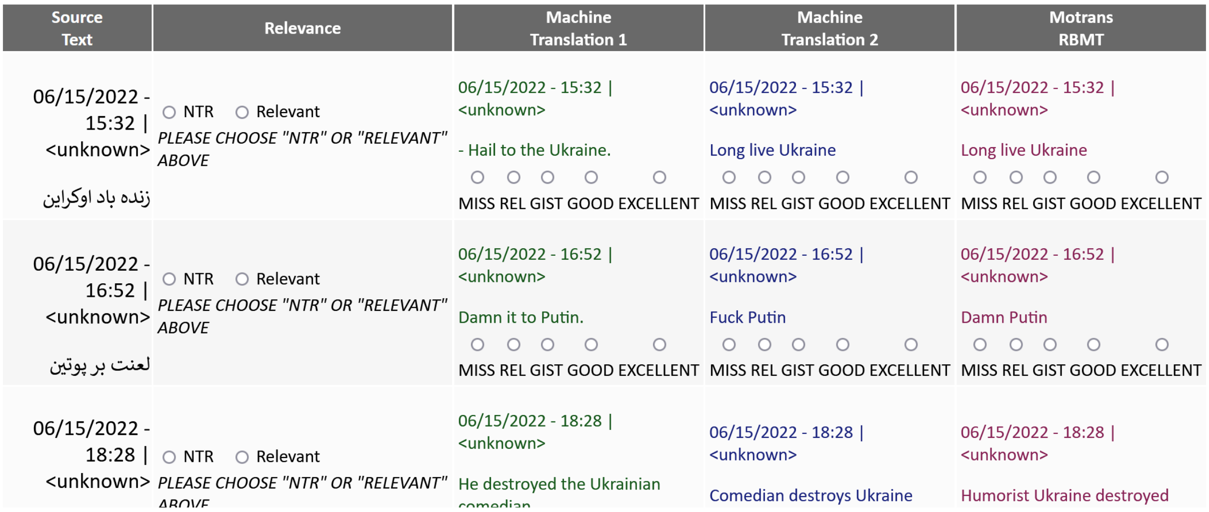
Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, New York, NY, USA. Association for Computing Machinery.

A Annotation Details

Screenshots of the annotation task are shown in Figures 5 and 6. Gold standard labels were assigned to items where both annotators agreed on the label. When annotators disagreed, the items were labeled ambiguous for the purpose of selecting items for the user study. Any ambiguous items that were eventually selected for the user study underwent a tie-breaking annotation where the original annotators were asked to come to an agreement on the final gold label. Interannotator agreement scores (Cohen's Kappa) before tie-breaking are shown in Table 9. Note that relevance applies to all items, but entity and sentiment apply only to items that both annotators labeled as relevant. Even before reconciliation, our annotators showed moderate to substantial agreement across the board and near-perfect agreement on entity and sentiment for Hezbollah and ISIS. The high level of agreement before reconciliation indicates that the annotators generally held the same understanding of the tasks and definitions, lending additional support to the reliability of the final reconciled labels.

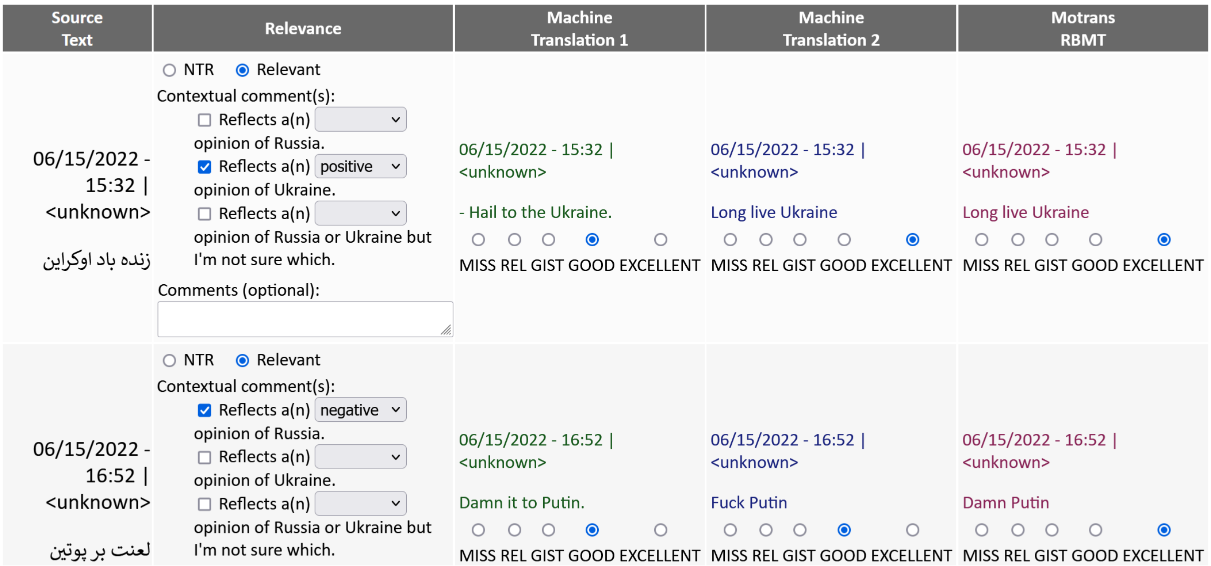
The distribution of relevance, entity, and sentiment labels for comments in each task is shown in Table 8. In total, 32 out of the 61 comments included in the user study were labeled relevant. Because the comments were selected in threads, the relevant entities are not evenly distributed between tasks. All of the relevant comments in Russia/Ukraine Task A relate to Russia, compared to only half of the relevant comments in Russia/Ukraine Task B. On the reverse, only one comment in Russia/Ukraine Task A relates to Ukraine compared to all but one comment in Russia/Ukraine Task B. The Hezbollah comments are more evenly split, with three in Hezbollah/ISIS Task A and two in Hezbollah/ISIS task B, but the ISIS-related comments are almost all in Task B, with only one in Task A. The Hezbollah/ISIS tasks also contain fewer relevant comments overall compared to the Russia/Ukraine tasks. This difference is likely due to the recency of Russia's war in Ukraine at the time the comments were collected.

The annotation also included a human evaluation. For each comment displayed in the thread context, the language analysts rated the outputs of NLLB-200, SYSTRAN, and Motrans using a task-focused adaptation of the evaluation scales from Licht et al. (2022). Each quality level was given a descriptive label to emphasize that they are not meant to be equally distanced points in a range but rather descriptive quality levels. The descriptions of Licht et al. (2022)'s levels 4-5 were retained, but the descriptions of the first three labels were adapted to fit the levels of comprehension in the user study task. The lowest quality label was *MISS*, described as a translation that is so different from the meaning of the source text that a non-language-enabled



Source Text	Relevance	Machine Translation 1	Machine Translation 2	Motrans RBMT
06/15/2022 - 15:32 | <unknown> زنده باد اوکراین	○ NTR ○ Relevant *PLEASE CHOOSE "NTR" OR "RELEVANT" ABOVE*	06/15/2022 - 15:32 | <unknown> - Hail to the Ukraine. ○ ○ ○ ○ ○ MISS REL GIST GOOD EXCELLENT	06/15/2022 - 15:32 | <unknown> Long live Ukraine ○ ○ ○ ○ ○ MISS REL GIST GOOD EXCELLENT	06/15/2022 - 15:32 | <unknown> Long live Ukraine ○ ○ ○ ○ ○ MISS REL GIST GOOD EXCELLENT
06/15/2022 - 16:52 | <unknown> لعنت بر پوتین	○ NTR ○ Relevant *PLEASE CHOOSE "NTR" OR "RELEVANT" ABOVE*	06/15/2022 - 16:52 | <unknown> Damn it to Putin. ○ ○ ○ ○ ○ MISS REL GIST GOOD EXCELLENT	06/15/2022 - 16:52 | <unknown> Fuck Putin ○ ○ ○ ○ ○ MISS REL GIST GOOD EXCELLENT	06/15/2022 - 16:52 | <unknown> Damn Putin ○ ○ ○ ○ ○ MISS REL GIST GOOD EXCELLENT
06/15/2022 - 18:28 | <unknown>	○ NTR ○ Relevant *PLEASE CHOOSE "NTR" OR "RELEVANT" ABOVE*	06/15/2022 - 18:28 | <unknown> He destroyed the Ukrainian comedian	06/15/2022 - 18:28 | <unknown> Comedian destroys Ukraine	06/15/2022 - 18:28 | <unknown> Humorist Ukraine destroyed

Figure 5: Screenshot of the relevance judgment and MT quality annotation view.



Source Text	Relevance	Machine Translation 1	Machine Translation 2	Motrans RBMT
06/15/2022 - 15:32 | <unknown> زنده باد اوکراین	○ NTR ● Relevant Contextual comment(s): ☐ Reflects a(n) opinion of Russia. ☑ Reflects a(n) positive opinion of Ukraine. ☐ Reflects a(n) opinion of Russia or Ukraine but I'm not sure which. Comments (optional):	06/15/2022 - 15:32 | <unknown> - Hail to the Ukraine. ○ ○ ○ ● ○ MISS REL GIST GOOD EXCELLENT	06/15/2022 - 15:32 | <unknown> Long live Ukraine ○ ○ ○ ○ ● MISS REL GIST GOOD EXCELLENT	06/15/2022 - 15:32 | <unknown> Long live Ukraine ○ ○ ○ ○ ● MISS REL GIST GOOD EXCELLENT
06/15/2022 - 16:52 | <unknown> لعنت بر پوتین	○ NTR ● Relevant Contextual comment(s): ☑ Reflects a(n) negative opinion of Russia. ☐ Reflects a(n) opinion of Ukraine. ☐ Reflects a(n) opinion of Russia or Ukraine but I'm not sure which.	06/15/2022 - 16:52 | <unknown> Damn it to Putin. ○ ○ ○ ● ○ MISS REL GIST GOOD EXCELLENT	06/15/2022 - 16:52 | <unknown> Fuck Putin ○ ○ ○ ● ○ MISS REL GIST GOOD EXCELLENT	06/15/2022 - 16:52 | <unknown> Damn Putin ○ ○ ○ ○ ● MISS REL GIST GOOD EXCELLENT

Figure 6: Screenshot of a completed comment annotation.

Task		Count	Relevant	Entity	Positive	Negative
Russia/ Ukraine	A	16	81.3% (13)	100% (13)	23.1% (3)	76.9% (10)
				7.7% (1)	0.0% (0)	100.0% (1)
Russia/ Ukraine	B	14	57.1% (8)	50.0% (4)	0.0% (4)	100.0% (4)
				87.5% (7)	57.1% (4)	42.9% (3)
Hezbollah/ ISIS	A	15	26.7% (4)	75.0% (3)	66.7% (2)	33.3% (1)
				25.0% (1)	0.0% (0)	100.0% (1)
Hezbollah/ ISIS	B	16	43.8% (7)	28.6% (2)	100.0% (2)	0.0% (0)
				71.4% (5)	0.0% (0)	100.0% (5)

Table 8: Distribution of gold standard relevance, entity, and sentiment labels for comments chosen for each task. *Relevant* indicates how often that entity was judged to be a relevant entity, and *Positive* and *Negative* indicate how often the sentiment towards that entity was positive or negative, respectively.

Label type	Russia/Ukraine	Hezbollah/ISIS	Combined
Relevance	0.537	0.566	0.574
Entity	0.684	0.834	0.740
Sentiment	0.679	0.972	0.799

Table 9: Annotator agreement (κ) on relevance, entity, and sentiment labels for our two annotators on the 556 comments (210 Russia/Ukraine; 346 Hezbollah/ISIS).

analyst would not be able to reliably make even a relevance judgment. The second level was *REL-ONLY*, described as translation quality sufficient to make a relevance judgment but with significant information missing or incorrect. The third level was *GIST*, described as translating critical information correctly but with less important information missing or incorrect. Levels 4 and 5 were labeled *GOOD* and *EXCELLENT* respectively. Translations below GIST quality (MISS or REL-ONLY) can be considered potentially misleading in this scenario.

Table 10 shows the percent of translations from each MT system that were given each of the labels and the percent that were potentially misleading (below GIST). Table 11 shows interannotator agreement (Kendall's Tau).

Motrans's lack of fluency is illustrated in the low percentage of translations that were at the GOOD or EXCELLENT level (9.53% and 5.94%, respectively), but its emphasis on adequacy is reflected in the smaller number of translations at the MISS level (10.79%) compared to NMT1 (17.45%) and NMT2 (13.13%). Because Motrans is rule-based MT, it cannot hallucinate or drop content as NMT models might, though it may mistranslate or leave words untranslated.

NMT2 (SYSTRAN) has the lowest percentage of Below GIST translations and the highest percentage of GOOD and EXCELLENT translations, suggesting that NMT2 might be a better match for these topics and this style than NMT1. However, these very specific domains (comments related to terrorist groups ISIS and Hezbollah and Russia's war in Ukraine) are only a small sample of domains that would need to be covered by a Persian-English MT system deployed to an intelligence analysis workforce. A multilingual model like NMT1 that demonstrates reasonable performance on a generic test set like FLORES may still be preferable as a baseline system, particularly if alternate NMT or RBMT proves beneficial in helping users overcome errors in the first NMT output.

MT	MISS	REL-ONLY	GIST	GOOD	EXCELLENT	Below GIST
NMT1	17.45%	31.12%	28.42%	13.13%	9.89%	48.56%
NMT2	13.13%	27.88%	32.19%	14.39%	12.41%	41.01%
Motrans	10.79%	38.67%	35.07%	9.53%	5.94%	49.46%

Table 10: Human quality judgments on all comment translations from NMT1 (NLLB-200), NMT2 (SYSTRAN), and Motrans.

Label type	Russia/Ukraine	Hezbollah/ISIS	Combined
NMT1	0.561	0.613	0.597
NMT2	0.666	0.561	0.602
RBMT	0.448	0.543	0.509
All	0.561	0.573	0.571

Table 11: Annotator agreement on MT quality labels (Kendall's tau) for our two annotators on the 556 comments (210 Russia/Ukraine; 346 Hezbollah/ISIS).