ABSTRACT

Title of Dissertation: WHEN GOOD MT GOES BAD: UNDERSTANDING AND MITIGATING MISLEADING MACHINE TRANSLATIONS

Marianna J. Martindale
Doctor of Philosophy, 2024

Dissertation Directed by: Dr. Marine Carpuat
Department of Computer Science

Machine Translation (MT) has long been viewed as a force multiplier, enabling monolingual users to assist in processing foreign language text. In ideal situations, Neural MT (NMT) provides unprecedented MT quality, potentially increasing productivity and user acceptance of the technology. However, outside of ideal circumstances, NMT introduces new types of errors that may be difficult for users who don't understand the source language to recognize, resulting in misleading output. This dissertation seeks to understand the prevalence, nature, and impact of potentially misleading output and whether a simple intervention can mitigate its effects on monolingual users.

To understand the prevalence of misleading MT output, we conduct a study to quantify the potential impact of output that is fluent but not adequate, or "fluently inadequate", by observing the relative frequency of these types of errors in two types of MT models, statistical and early neural models. We find that neural models were consistently more prone to this type of error than traditional statistical models.

However, improving the overall quality of the MT system such as through domain adaptation reduces these errors.

We examine the nature of misleading MT output by moving from an intrinsic feature (fluency) to a more user-centered feature, believability, defined as a monolingual user's perception of the likelihood that the meaning of the MT output matches the meaning of the input, without understanding the source. We find that fluency accounts for most believability judgments, but semantic features like plausibility also play a role.

Finally, we turn to mitigating the impacts of potentially misleading NMT output. We propose two simple interventions to help users more effectively handle inadequate output: providing output from a second NMT system and providing output from a rule-based MT (RBMT) system. We test these interventions for one use case with a user study designed to mimic typical intelligence analysis triage workflows and with actual intelligence analysts as participants. We see significant increases in performance on relevance judgment tasks with output from two NMT systems and in performance on relevant entity identification tasks with the addition of RBMT output.

WHEN GOOD MT GOES BAD: UNDERSTANDING AND MITIGATING MISLEADING MACHINE TRANSLATIONS

by

Marianna J. Martindale

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Professor Marine Carpuat, Chair/Advisor
Professor Ge Gao
Professor Hernisa Kacorri
Professor Susannah Paletz
Professor Philip Resnik (Dean's Representative)

Dedication

In memory of Dr. Dave and the CAMT:

Ita bonum est ut muneri rei publicae satisfaciat.

Acknowledgments

My PhD experience has been a journey with more guides along the way than I can even begin to thank by name, not only because there are so many but because some of their names would be redacted for national security reasons. I could not have made it without my academic and professional peers and mentors, as well as the friends and family who have offered their love and support and far more optimism than I ever managed to muster for myself. I may have occasionally resented you for telling me that I could do it when I was sure I couldn't, but you were right all along.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACTFL	American Council on the Teaching of Foreign Languages
AD	Adequacy
AI	Artificial Intelligence
AIS	Analytic Integrity Standards
ARPA	Advanced Research Projects Agency
BL	Believability
BLEU	BiLingual Evaluation Understudy
BPE	Byte Pair Encoding
BVSS	Bag-of-Vectors Sentence Similarity
CLIR	Cross-Language Information Retrieval
DA	Direct Assessment
FL	Fluency
HLTCOE	Human Language Technology Center of Excellence
IA	Intelligence Analyst
ICD	Intelligence Community Directive
IEEE	Institute of Electrical and Electronics Engineers
IRB	Institutional review Board
IWSLT	International Conference on Spoken Language Translation
LSTM	Long Short-Term Memory
MT	Machine Translation
MTTT	Multi-Target TED Talks Task
NLP	Natural Language Processing
NMT	Neural Machine Translation
RBMT	Rule-Based Machine Translation
SMT	Statistical Machine Translation
SNLI	Stanford Natural Language Inference (Corpus)
STS	Semantic Textual Similarity
TER	Translation Edit Rate
UMD	University of Maryland
UN	UNited Nations
USCIS	United States Citizenship and Immigration Services
WMT	Workshop on Machine Translation

Chapter 1: Introduction

1.1 Motivation

Machine Translation (MT) is a technology that takes as its input text in one human language (the *source* language) and produces text in another human language (the *target* language) that is expected to convey the meaning of the source text. Initial attempts at building MT relied on dictionaries and rules to transform text between languages (Rule-Based MT, RBMT), producing consistent but often disfluent translations that may be unreadable without human post-editing (Bar-Hillel, 1951), to mixed reviews from users (Yang and Lange, 1998). Statistical machine translation (SMT) approaches improved fluency and allowed automatic training of broad coverage MT systems from aligned previously translated text, leading to widespread availability through companies like Google, whose Translate service processed more than 100 billion words per day from more than 500 million users in 2016 (Turovsky, 2016). Neural (NMT) approaches have provided further improvements, sometimes scoring as well as human translators in sentence-level evaluations on text from the domain the model was trained on (Hassan Awadalla et al., 2018; Barrault et al., 2019, 2020), although the notion that these results represent "human parity" has been widely refuted (Läubli et al., 2018; Toral et al.,

2018; Läubli et al., 2020; Toral, 2020; Graham et al., 2020a,b). Despite these quality improvements, MT remains imperfect. Those imperfections can have consequences including embarrassment (Cuthbertson, 2020; Liebling et al., 2020), potential health and safety risks (Liebling et al., 2020, 2021; Torbati, 2019), and even one infamous incident where a man was taken into custody based on incorrect MT of a social media post (Berger, 2017).

With MT's near-ubiquitous availability, it is crucial to understand the circumstances in which it may mislead users. The goal of this dissertation is to better understand the nature of potentially misleading MT output, how users react to these errors, and whether implementing simple interventions can mitigate risks and help users more accurately calibrate their confidence in judgments they make based on MT output in a scenario that mimics a real-world use case.

A user can be said to have been misled by MT output when they incorrectly judge its accuracy, particularly when they accept and act on incorrect translations and, to a lesser degree, when they reject accurate information. The likelihood of being misled and the impact of these misperceptions depend on the user and the use case.

There are many use cases for MT, and these use cases can broadly be categorized as MT for dissemination, assimilation, and communication (Koehn, 2010). In dissemination use cases, MT is used in the human translation process to improve translator efficiency and sometimes quality. The MT user is producing a translation to share information with an audience. Without MT, the content would be translated less efficiently, or perhaps the additional cost of translation without MT would

result in the content not being shared with that audience. Because the user knows both the source and target language, they are less likely to be misled by incorrect MT output. In assimilation use cases, the user is an information consumer who wants to glean information from text written in a language they do not understand. In these use cases, the content generally would not be translated without MT, and the user would simply not have access to the information. These use cases can include casual browsing on the web and social media as well as formal uses such as academic research or intelligence analysis. Assimilation users are more likely to be misled than dissemination users because they do not know the source language, but the impact depends on the specific use case, as discussed below. In both dissemination and assimilation use cases, the source material is generally static, and the MT user is either the provider or consumer of translated information. Communication use cases, on the other hand, include elements of assimilation and dissemination, with participants both producing and consuming dynamic content interactively with opportunities to rephrase or request clarification when the MT output is unclear. Because participants are not fluent in the same language, there is more chance of being misled than in dissemination use cases, but there is also more opportunity to recover through clarification and negotiation than in assimilation use cases.

We have chosen to focus on an assimilation use case in this work because of its risk profile. In assimilation use cases, the user has limited ability to recognize errors in MT output because they do not understand the source language, and there is no opportunity to request clarification from a static source text. As such, it is particularly important to understand not only how often MT errors occur but also

how users respond to them. If users are able to recognize and reject MT errors, the potential negative impacts will only affect the user's acceptance of the MT system. However, if users do not recognize errors in meaning, we can say that the user is misled by the MT output. Similarly, if users reject correct information in MT output based on a belief that the MT is not accurate, the user has also been misled. The potential negative impacts of either type of misleading MT output will depend on the specific use case.

The use case we study in this work is MT-enabled scanning, in which intelligence analysts assist in triaging foreign language documents to identify those that are worth translation and further analysis. By studying this use case, we seek to learn how potentially misleading MT output affects analyst performance and confidence and whether simple interventions can improve the analyst experience and effectiveness in using MT. These observations can serve to inform related use cases and future studies on MT misleadingness.

1.2 Roadmap

In this dissertation, we first seek to understand misleading machine translations by reviewing the relevant literature and conducting studies related to aspects of MT misleadingness. We then explore potential mitigations for misleading MT output through a user study on MT-enabled scanning for intelligence analysis.

Chapter 2 establishes a foundation for the remainder of the work by reviewing relevant literature on assessing the misleadingness of MT output and conducting

user studies. We begin with definitions and theory related to credibility and trust. We then review standard MT quality evaluation practices and results of evaluations related to misleadingness.

Next, we introduce work related to the characteristics of potentially misleading MT output in Chapters 3 and 4, first exploring potentially misleading MT output through the lens of common MT evaluation features, *fluency* and *adequacy*, and then introducing the idea of MT *believability*.

Chapter 3 reviews a pilot study that measured the effects of fluency and adequacy errors on user trust in MT at the system level and the individual translation level and then details an experiment to quantify the frequency of fluency and adequacy errors in statistical and neural MT outputs. The pilot study results suggest that translations that make adequacy errors fluently may mislead users. Given the impressive fluency of modern MT systems, which may include output that is fluent but not adequate, or (*fluently inadequate*), the experiment we conducted sought to identify those errors and quantify their frequency in MT output of varying quality (Section 3.2). To that end, we introduced a method for automatically predicting whether translated segments are fluently inadequate by predicting fluency using grammaticality scores and predicting adequacy by augmenting sentence BLEU with a novel Bag-of-Vectors Sentence Similarity (BVSS). We then applied this technique to analyze the outputs of statistical and neural systems for six language pairs with different levels of translation quality. We found that neural models were consistently more prone to this type of error than traditional statistical models. However, improving the overall quality of the MT system, such as through domain adaptation,

reduced these errors.

Turning to the question of how users respond to potentially misleading MT output, Chapter 4 explicitly tests whether fluently inadequate translations would actually mislead users by annotating MT outputs for fluency, adequacy, and believability. Users who do not understand the source language respond to MT output based on their perception of the likelihood that the meaning of the MT output matches the meaning of the source text. We refer to this as *believability*. Output that is not believable may be off-putting to users, but believable MT output with incorrect meaning may mislead them. In this study, we examined the relationship of believability to fluency and adequacy by applying traditional MT direct assessment protocols to annotate all three features on the output of neural MT systems. Quantitative analysis of these annotations shows that believability is closely related to but distinct from fluency, and initial qualitative analysis suggests that semantic features may account for the difference.

In Chapter 5, we move from understanding misleading MT to mitigating it. We conduct a user study to determine how often potentially misleading MT outputs mislead users and to explore practical interventions that may help mitigate the risks of misleading MT. Having observed characteristics of potentially misleading translations in a general sense in Chapters 3 and 4, the question remains of how these translations affect users in the context of a real-world use case. The use case we study is MT-enabled scanning of foreign language documents by intelligence analysts. We recruit actual intelligence analysts as participants and ask them to perform tasks similar to typical scanning tasks based on MT output. We measure

their performance and confidence under variations of three conditions: with only one MT output, with outputs from two neural MT systems, and with outputs from a rule-based MT system. This is the publicly available first study to address the use of MT in intelligence analysis and the first to explore the use of multiple MT outputs to mitigate misleading MT output.

We conclude by summarizing the contributions of this work along with its limitations and potential topics for future work in Chapter 6.

Chapter 2: Background

This chapter explores prior research related to the work in this dissertation, drawing from the fields of communications, human-computer interaction (HCI), artificial intelligence (AI), computational linguistics, and translation studies. First, we seek to establish the theoretical and definitional foundation for this work by exploring concepts related to misleadingness, namely, computer credibility and trust in automation. These ideas are primarily addressed in communications, HCI, and AI work. We then look to prior work in MT evaluation from computational linguistics and translation studies to identify the types of errors that may relate to misleading MT output and potential approaches to identifying misleading MT. Finally, although there have been no previously published user studies on the use of MT for intelligence analysis, there have been user studies on machine translation in other use cases and addressing similar research questions with other technologies. We discuss several of these studies from the HCI, translation studies, and MT literature to identify similarities to this work and gaps that this work intends to fill.

2.1 Definitions and Theory

Recall that there are two ways MT output can mislead a user: (1) the MT output does not reflect the meaning of the source text while the user believes that it does, and (2) the MT output does reflect the meaning of the source text, but the user believes that it does not. Both represent an inaccurate perception of the *believability* or *credibility* of the MT and/or inaccurately calibrated *trust* in the MT system. Exploring the credibility literature can help us understand these incidents more fully.

In their seminal work on computer credibility, Fogg and Tseng (1999) bring together prior research on human-to-human credibility and early work on computer credibility and propose conceptual frameworks around types of computer credibility, types of errors in credibility judgment, and models of credibility judgment. They define credibility as synonymous with believability and note that users demonstrate their perception of credibility when they, "trust the information," "accept the advice," or "believe the output" from a system (Fogg and Tseng, 1999, p. 81). The types of misleading MT align with the types of errors users make in judging credibility: the false positive *gullibility error* (the user believes inaccurate output) and the false negative *incredulity error* (the user rejects accurate output).

In the broader credibility literature, three high-level elements are involved in credibility judgments: source credibility, media credibility, and message credibility (Metzger et al., 2003; Rieh and Danielson, 2007). Source credibility relates to the provider of the content. In the case of MT, the source would be the perceived

provider of the MT. This could be the creator of the MT model if the user is aware of that information and/or the provider of the interface through which the user accesses the MT (e.g., AltaVista's "Babelfish" translation service used SYSTRAN as its MT provider (Yang and Lange, 1998)). Media credibility relates to the perception of the means by which the content is conveyed. Traditionally, this meant baseline perceptions of broad categories of media like print, radio, and television (Metzger et al., 2003), but within the domain of web credibility, has been expanded to include characteristics of the site on which content is presented, such as the overall look and feel (Kim, 2010). Analogous to the traditional definition of media credibility would be a user's baseline perception of MT as a technology. The interface through which a user accesses MT would be the equivalent of the website feature elements of media credibility. Message credibility refers to content features, such as reasonableness (Liu, 2004; Kim and Oh, 2009; Kim, 2010; John et al., 2011) and spelling and grammar (Fogg et al., 2001; Everard and Galleta, 2005; Metzger et al., 2010; Chesney and Su, 2010; John et al., 2011). These features are directly applicable to MT output and, as we will see in Section 2.2, are also related to MT evaluation.

Fogg (2003) proposes *Prominence-Interpretation Theory*, a theory of how users bring together features of online material to make credibility judgments of the types discussed above. A feature's impact on the user's credibility judgment is based on the *prominence* of the feature to the user (i.e., the extent to which they are aware of it) and the user's *interpretation* or judgment of the feature. Elements that may affect the prominence of a feature include user involvement, the topic of the content, the user's task or goal in accessing the content, the user's prior

experience, and individual user characteristics. Once the user is aware of a feature, their interpretation of it will include factors such as their assumptions, their skills or knowledge, and the context in which they encounter it. This suggests that an information provider can affect the user's credibility judgments by manipulating the prominence of features or influencing the user's interpretation of them.

Closely related to credibility is the idea of *trust*. Fogg and Tseng (1999) suggest that credibility and trust are often conflated, but trust is tied to the system rather than a particular output and more aptly likened to reliability or dependability rather than believability. Consistent with this observation, trust in automation and trust in AI studies have been focused on systems that directly perform tasks as in industrial automation (e.g., Lee and Moray, 1994) or recommend actions as in route planning (e.g., de Vries et al., 2003) or memory recognition (Yang et al., 2016). In dissemination use cases, MT performs a similar role: it directly helps the user accomplish the task of translation. Trust has been measured in several ways, including simply asking the user (Yang et al., 2016) or measuring the user's expectations of system performance and actions taken (de Vries et al., 2003). In the context of MT, (Cadwell et al., 2018) noted translators' lack of trust in MT based on comments recorded during focus groups.

However, in assimilation and communication use cases, the user's end goal is not translation: MT is merely a transformative process, providing information to be used in accomplishing the end goal. Jacovi et al. (2021) apply principles of human-human trust to trust in AI and conclude that the concept of human-AI trust is best applied to systems the user attributes *intent* to as opposed to systems perceived as

inanimate, where the relationship is one of reliance. If the user does not perceive the system as an agent, they do not assess trust in the system and merely rely on the output. In the same work, Jacovi et al. also posit that to have trust, the user must perceive and accept the risk of negative consequences from relying on inaccurate output from the model. When these factors come together, users can respond to model errors similarly to how they respond to betrayal in human relationships. Yang et al. (2016) measured user trust in an automated decision aid and found that changes in trust did not follow a purely statistical process where incorrect suggestions would have a consistently negative effect on trust. Instead, incorrect suggestions had a stronger negative effect when the user believed and acted on the suggestion than when they went against it. This suggests that if MT users are aware of the MT system, see it as an agent in the translation process, choose to rely on misleading output, and subsequently learn that the MT was inaccurate, they may have a strongly negative reaction and dramatically reduce their trust in the system. If, on the other hand, the user is minimally cognizant of the MT system, they will primarily make credibility judgments of particular outputs rather than modulating their trust in the system.

Jacovi et al. (2021)'s causes and types of trust in AI share some similarities with the types and features of credibility. *Intrinsic trust* is trust based on the user's observations of exposed features of the model's reasoning process. The key elements of intrinsic trust are that the user understands the model's reasoning process and that the reasoning process is in accord with the user's expectations and assumptions. Note that the first element is similar to prominence, and the second is interpretation.

Extrinsic trust, according to the Jacovi et al. framework, is based on observations of the model outputs, or in other words, evaluation. Users may be more likely to trust an MT system if they are made aware of evaluations where the system performed well.

2.2 MT Quality Evaluation

The two factors of MT misleadingness are whether or not the meaning of the MT output matches the meaning of the source text and whether or not the user believes that it does. In this section, we discuss frameworks for MT evaluation that may relate to misleadingness and results of prior MT evaluation studies that may shed light on the prevalence of potentially misleading MT output.

2.2.1 Frameworks for MT Quality Evaluation

MT quality evaluation seeks to objectively judge how well MT performs the task of translation. These evaluations allow MT researchers to measure progress and compare systems to each other and to human translator performance.

In the 1990s, the Advanced Research Projects Agency (ARPA) funded an initiative to improve MT. To measure that improvement, it was necessary to establish evaluation criteria (White et al., 1994). They used the same high-level features used in most evaluation efforts since then: *fluency* and *adequacy*. Fluency was defined as a judgment of "whether the translation reads like good English ... without knowing the accuracy of the content," and adequacy, an assessment of "the degree to which

the information in a professional translation can be found in an MT (or control) output of the same text" (White et al., 1994, p. 196).

Others have adopted fine-grained error categories (e.g., Flanagan, 1994; Vilar et al., 2006; Lommel et al., 2014) to enable more detailed error analysis. Such frameworks have been used in comparing human post-editing effort for different error types (Temnikova, 2010; Kirchhoff et al., 2014; Koponen, 2012) and in comparative error analysis between statistical (SMT) and neural (NMT) paradigms (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Klubička et al., 2017).

The Multidimensional Quality Metrics (MQM) framework[1] merges the coarse-grained features of fluency and adequacy (accuracy in MQM) with fine-grained error categories in a hierarchical approach intended to standardize evaluation of both human and machine translation and allow those applying the framework to distinguish between error types that are more or less significant in a specific application (Lommel et al., 2014). Because this framing is particularly useful for dissemination use cases, in addition to fluency and accuracy, they also include features primarily useful for publication, such as design and locale conventions (e.g., date formats). Accuracy errors in MQM include additions, omissions, untranslated, and mistranslated, which can be further specified based on the type of mistranslated content (e.g., name, number) or the type of translation error (e.g., introduced ambiguity, false friend). Fluency errors include ambiguity, coherence, spelling, and grammatical errors, which can be further subdivided on two additional levels.

How do these error types relate to misleadingness? Recall that the two di-

[1] http://qt21.eu/mqm-definition

mensions of misleadingness are whether the MT accurately reflects the meaning of the source text and whether the user believes that it does. The message credibility features mentioned in Section 2.1 include reasonableness and spelling and grammar errors, which are both related to fluency. The fluency features also share the characteristic of being able to be evaluated without knowledge of the source language, which is the precise scenario in which MT output may be misleading.

2.2.2 Results of MT Quality Evaluations Related to Misleadingness

Evaluations that assess both fluency and adequacy can suggest potentially misleading MT output. The annual Conference (formerly Workshop) on Machine Translation (WMT) is one large-scale, ongoing MT quality evaluation. The 2006 and 2007 workshops included direct assessment of fluency and adequacy Koehn and Monz (2006); Callison-Burch et al. (2007), but the inter-annotator agreement was low, and the high correlation between fluency and adequacy annotations suggested that the annotators may have had difficulty distinguishing the features (Callison-Burch et al., 2007). In 2016, direct assessment of fluency and adequacy was reintroduced (Bojar et al., 2016), but it was determined that adequacy was the more important measure so fluency direct assessment was dropped from 2017 on Bojar et al. (2017, 2018); Barrault et al. (2019, 2020). Unfortunately, this change coincided with the widespread switch to NMT, so there are few fluency and adequacy annotations for NMT systems to assess the frequency of potentially misleading output based on these features. The study in Section 3.2 seeks to fill this gap by applying automatic

fluency and adequacy metrics to output from multiple SMT and NMT systems.

Several other studies have performed error analyses comparing NMT to SMT, showing that NMT produces output that is more fluent than SMT (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Koehn and Knowles, 2017). Koehn and Knowles (2017) specifically note the propensity of NMT to sometimes produce output that is low adequacy and even unrelated to the input–particularly when not trained on sufficient in-domain data. This problem has been referred to as *hallucination* in, e.g., Raunak et al. (2021); Wang and Sennrich (2020); Müller et al. (2020). The definitions of hallucination in each of these papers involve MT output that is not adequate but is at least somewhat fluent, which meets the characteristics of potentially misleading MT output described above, but their research focuses on identifying, explaining, and preventing hallucinations, and does not involve actual users to determine whether they are misled.

The analyses in Popović (2020a,b) get one step closer to misleadingness by focusing on *comprehensibility*. A translation is comprehensible if the reader can understand it (Popović, 2020b). However, Popović asserts that inadequate but comprehensible translations are misleading but does not test this assertion. It can be assumed that incomprehensible translations will not be misleading because the user will not be able to glean incorrect information from the translation if they cannot glean any information. However, whether the user *believes* the translation is adequate and would therefore be misled cannot be determined without asking them. Thus, comprehensibility is necessary but may not be sufficient for determining misleadingness.

As we have seen, MT quality assessment can be useful for identifying features of MT output with the *potential* to mislead, but determining whether MT is misleading requires explicitly addressing whether users believe the output of the MT system. The study in Chapter 4 seeks to fill this gap by collecting and analyzing believability annotations.

2.2.3 Automatic MT Quality Evaluation and Estimation

The most common type of MT evaluation is automatic MT evaluation. The advent of automatic MT evaluation metrics including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), and more recently chrF (Popović, 2015) and COMET (Rei et al., 2020) has allowed the field to develop and test models rapidly without costly human evaluations. These metrics rely on human translations as references to compare against and are particularly useful for ranking systems to observe overall improvements in quality on shared test sets. They cannot tell a user whether or not to trust the output for specific input text.

Referenceless evaluation comes in the form of confidence scores derived from the model itself or quality estimates based only on the input and output (Specia et al., 2009). Confidence scores can be fairly straightforward for statistical MT (e.g., Ueffing et al., 2003) and naive confidence scores are an inherent by-product of the NMT process, though these 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). Because translation is a complex task with many

types of errors, even well-calibrated scores have limited utility for users if they do not provide enough information to help the user recognize what went wrong. Well-calibrated word-level confidence can localize potential errors, but the severity of those issues is unclear. The Fine-Grained Error Detection subtask of the WMT Shared Task on Quality Estimation (Blain et al., 2023) introduced in 2023 fills this gap by evaluating models for identifying erroneous spans of MT output and labeling them according to the severity of the error. Although the ideas are promising, they are not yet enough to counter misleading MT output as the models so far have low precision and recall, with an F1 score of only 0.284 for the best-performing model for any language pair.

2.3 Assessing MT in Context

MT quality evaluation occurs in a vacuum: it tells us how well the MT performs, but not whether it is *useful*. To measure the utility of MT and allow documentation of the effects of potentially misleading MT output, MT must be assessed in the context of a task or use case. Prior work in this area includes task-based MT evaluation and MT user studies.

2.3.1 Task-Based MT Evaluation

Task-based MT evaluation is a way to judge the utility of MT for a particular type of use. The FEMTI framework approaches MT evaluation from the perspective of first considering the use case and the needs of that use case to identify appropriate

evaluation methods (King et al., 2003). Other task-based evaluation approaches focus on the comprehension aspect of MT use rather than specific use cases (e.g., Voss and Tate, 2006; Jones et al., 2007; Forcada et al., 2018). Similar to Popović's comprehensibility annotations, comprehension is only the first step of MT use, and without the context of applying the user's understanding to a particular problem, there is room for ambiguity on whether or not the user would actually be misled.

Resnik (1997) and Oard and Resnik (1999) provide a model for developing experimental tasks to evaluate the utility of MT on real-world tasks where users make a decision based on the MT output. The experimental task should be analogous to the real-world task and practical to create and evaluate, but abstracting the real-world task to the experimental task should not introduce biases that would not be present in the true task. The example used in the study is a card-sorting task based on the real-world use case of web browsing. Participants were asked to sort cards representing businesses from the Nihongo Yellow Pages into categories based on either human translations of the Japanese information from the corresponding entry in the Yellow Pages or "gists" of the Japanese information created using a simple dictionary lookup approach. Note that this task requires a similar level of comprehension to the relevance judgment aspect of MT-enabled scanning. Participants performed almost as well with the dictionary lookup as with the human translations. Although it was a small study, the results provide reason to suspect that if even minimal "translation" is enough to provide surprisingly effective topic identification, the task of relevance judgment may not be complex enough to observe the effects of potentially misleading MT output. On the other hand, it also opens

up the possibility that even a simple dictionary lookup tool may provide a useful signal in helping users interpret potentially misleading output.

2.3.2 MT User Studies

There have also been user studies focused on observing how MT is used rather than evaluating MT. Although none of these studies have addressed intelligence analysis use cases, some have addressed similar use cases or research questions.

The most closely related user studies to the use case we wish to address are other MT use cases where the MT is used primarily to make a relevance or importance judgment. MT use cases related to patent translation may be the most similar use cases to mine. Rossi and Wiggins (2013), Tinsley (2017), and Nurminen (2019) point to scanning as a primary use of patent MT. Nurminen (2019) surveyed patent professionals to learn how they use MT and what factors they use in assessing risk. The decisions made based on MT include identifying a patent as relevant prior art, choosing patent applications to monitor, and even identifying patents used in patentability decisions. Taking the next step of obtaining a human translation depends on the associated costs of translation and the risks of making incorrect decisions based solely on the MT output. Patent professionals reduce the risk of relying on MT by using contextual information such as their background knowledge and drawings and formulas to validate MT output. The risk is also reduced by the iterative nature of patent actions, allowing opportunities to correct mistakes. Providing additional contextual information and including opportunities for review

may also reduce risk from MT in intelligence analysis workflows.

The use cases in other studies are not as similar to the intelligence analysis use case, but they still provide valuable insights when they explore related research questions, particularly those that address interventions that provide additional information alongside an MT output. Gao et al. (2013) looked at the effect of adding keyword highlights to MT output in a communication setting. This intervention increases the prominence of the keywords, helping users to focus on important information and be less distracted by fluency errors. The additional information that Xu et al. (2014) and Gao et al. (2015) provide is the output from a second MT system. In both cases, the MT systems were SMT. Although the use case differs from ours, the intervention is the same. In the first study (Xu et al., 2014), users participated in a casual MT-mediated conversation and then were surveyed and interviewed about their experience. Participants preferred the two-MT scenario and cited differences between the outputs as particularly helpful. This matches the intuition that differences between MT outputs add prominence to those elements, allowing the user to consciously consider whether to believe the MT output. Similar to the first experiment, users in the second experiment (Gao et al., 2015) participated in an MT-mediated conversation and were surveyed about their experience, but instead of casual conversation, it was a task-oriented conversation. Performance on the navigation task was tracked, and the conversation transcript was annotated. In the two-MT condition, the teams performed better and reported less difficulty understanding. This suggests that in this communication use case, users are able to use the MT output more effectively when they can compare two outputs. Key

differences between these use cases and our study are the interactive nature of the task and the use of NMT rather than SMT. It seems likely that analysts will be able to learn from differences between MT outputs despite not being able to ask clarification questions, but it remains to be seen whether the increased fluency of NMT output and NMT-specific errors such as hallucination will be overcome.

Many MT user studies rely on surveys to understand the use cases, benefits, and shortfalls, including Yang and Lange (1998); Smith (2003); Nurminen and Papula (2018); Nurminen (2019); Liebling et al. (2020, 2021). Although a survey would not be specific enough to diagnose the circumstances in which MT is misleading for our use case, the results of prior surveys can provide background on the potential impact of misleading translations (Nurminen, 2019; Liebling et al., 2020, 2021). In both Liebling et al. (2020, 2021), users were frustrated by MT errors. The strategies to recognize errors observed in Liebling et al. (2021) included the use of external sources such as another MT tool or online dictionaries. This willingness of MT users to look to additional sources may also occur with analysts in the MT-enabled scanning use case.

2.4 Summary

In this chapter, we have reviewed prior work across diverse fields, including not only computational linguistics and translation studies–traditional venues for MT research–but also HCI, AI, and communications to identify relevant concepts and experimental results that inform our work. This work seeks to fill the gaps at

the intersection of credibility and MT evaluation and apply relevant concepts to the under-studied area of assimilation use cases, focusing specifically on the MT-enabled scanning use case.

Chapter 3: Understanding MT Misleadingness: Fluency

Users of Machine Translation who do not understand the source language have limited means of judging the accuracy of MT output. The primary features of standard evaluation of MT systems are fluency and adequacy. Fluency has been defined as a judgment of "whether the translation reads like good English...without knowing the accuracy of the content," and is typically combined with *adequacy*, an assessment of "the degree to which the information in a professional translation can be found in an MT (or control) output of the same text" (White et al., 1994). Of these features, only fluency can be judged monolingually. It is thus reasonable to assume that highly fluent MT outputs that are inadequate will be misleading. We previously conducted a pilot study (Martindale and Carpuat, 2018) to validate this assumption by observing the effects of fluency and adequacy in user trust in MT, summarized in section 3.1. Building on those results, we sought to quantify the potential impact of output that is fluent but not adequate, or "fluently inadequate", by observing the relative frequency of these categories of errors in two types of MT models, statistical and early neural models. This study is detailed in Section 3.2.

3.1 Summary of Pilot Study

The pilot study introduced a method of studying how users modulate their trust in an MT system after seeing errorful (disfluent or inadequate) output amidst good (fluent and adequate) output. Prior work outside of MT, such as Yang et al. (2016), found that users who accepted a system output and then learned that the output was incorrect lowered their trust in the system more than users who did not accept an incorrect output. If their findings also apply to the machine translation scenario, we would expect that user trust in the MT system will go down further when they see this type of bad translation compared to merely disfluent translations. However, participants in the pilot study showed no statistically significant reduction in trust after learning that a fluent translation was not adequate, reacting more strongly to fluency errors than to adequacy errors. Although only a small pilot study, this finding suggests that fluency plays a powerful role in user trust and may cause users to be misled when a translation is inadequate.

3.2 Identifying Fluently Inadequate MT Output

If users are misled by MT output that is fluent but not adequate, it is important to understand how often these translations occur.

3.2.1 Introduction

Prior work showed that well-trained, in-domain neural machine translation (NMT) systems can produce translations that, at the sentence level, are rated on par with human reference translations (Hassan Awadalla et al., 2018). Part of this success comes from the impressive improvements in the fluency of NMT output compared to previous MT paradigms (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Koehn and Knowles, 2017). However, NMT has also been shown to sometimes produce low adequacy output and even unrelated to the input–particularly when not trained on sufficient in-domain data (Koehn and Knowles, 2017). Because of NMT's ability to produce fluent output, these translations may not just be inadequate but *fluently inadequate*. The fluency of *fluently inadequate* translations may mislead users into trusting the content based on fluency alone–particularly in the context of other fluent and adequate translations (Martindale and Carpuat, 2018).

Mitigating the effects of fluently inadequate translations first requires understanding the scale of the problem and what situations are likely to generate these errors. The general success and high system-level quality of NMT suggest that fluently inadequate translations are rare, but we cannot say how rare without a means of automatically identifying *potentially* fluently inadequate translations in large collections of MT output. In this work, we propose a method to automatically detect fluently inadequate translations based on the underlying characteristics of fluency and adequacy. We view fluently inadequate translations as translations that are

fluent, well-formed sentences that could have been written by a human and that do not preserve the meaning of the reference. In practice, given a reference translation r and MT hypothesis h, we consider h to be fluently inadequate if $\mathit{fluency}(h) > \tau_f$ and $\mathit{adequacy}(h, r) < \tau_a$, where τ_a and τ_f are minimum fluency and adequacy thresholds respectively. We define novel fluency and adequacy metrics for this purpose, building on prior work on grammaticality detection and comparisons of multisets applied to word embeddings.

We conduct two sets of experiments. First, we evaluate the fluency and adequacy metrics, establishing that they can be used to detect fluently inadequate translations, and set thresholds τ_a and τ_f empirically in Section 3.2.4. We then conduct an automatic analysis to assess how frequent these errors are in neural and statistical machine translation (SMT) systems for a variety of languages and varying levels of model quality and train/test domain match. We find that fluently inadequate translations are more common in NMT overall, especially when there is less training data and when there is a mismatch between training and test data.

3.2.2 Predicting Fluency

We propose to score fluency using metrics introduced for the related task of detecting grammaticality, which scores a sentence's well-formedness. Although grammaticality focuses on well-formedness while fluency includes all aspects of "sounding natural," the metrics used to predict grammaticality may still prove to be good measures of fluency. Lau et al. (2017) take an unsupervised language modeling ap-

proach to predicting grammaticality. Based on the intuition that well-formedness errors will be caused by one or more incorrect or out-of-place words, they introduce scores based not only on sentence probability but also on scores that focus on the lowest word probabilities in a segment. Specifically, given a 5-gram language model, the following scores are computed:

$$Mean\ LP = \frac{\sum_{n=1:N} \log p_5(w_n|w_{n-1}, ...)}{N} \tag{3.1}$$

$$Norm\ LP = \frac{\sum_{n=1:N} \log p_5(w_n|w_{n-1}, ...)}{\sum_{n=1:N} \log p_1(w_n)} \tag{3.2}$$

$$Word\ LPmin_n = min_n \left\{ -\frac{\log p_5(w)}{\log p_1(w)} \right\} \tag{3.3}$$

$$Word\ LP_{n\%} = \frac{\sum_{w \in LP_{n\%}} -\frac{\log p_5(w)}{\log p_1(w)}}{|LP_{n\%}|} \tag{3.4}$$

$$Word\ LP_{mean} = \frac{\sum_{n=1:N} -\frac{\log p_5(w_n)}{\log p_1(w_n)}}{N} \tag{3.5}$$

where $logp_5$ is the 5-gram log probability, w_n is the n^{th} word and N is the number of words in the sentence. $Mean\ LP$ is the sentence n-gram log probability, normalized by length. *NormLP* is the sentence n-gram log probability, normalized by sentence unigram log probability. The other metrics are focused on the probability of individual words given the preceding words. Each word's 5-gram log probability

is normalized by its unigram probability ($logp_1$), $Word\ LPmin_n$ is the n^{th} lowest normalized word probability, $LP_{n\%}$ is the lowest $n\%$ normalized word probabilities. Because there may be outliers that score artificially low, we introduce an additional variant, $WordLP_{mid}$, which uses LP_{mid}, the middle 50% of the normalized word probabilities:

$$Word\ LP_{mid} = \frac{\sum_{w \in LP_{mid}} -\frac{\log p_5(w)}{\log p_1(w)}}{|LP_{mid}|} \tag{3.6}$$

We expect that fluently inadequate output is being influenced by the training data more than the input text, so we build our language model based on the target side of the system training data rather than a large generic language model.

3.2.3 Predicting Adequacy

Prior work on predicting adequacy includes Semantic Textual Similarity (STS) and automatic MT evaluation. Adequacy is, essentially, semantic equivalence, and SemEval's STS task aims to measure the degree of semantic equivalence between two sentences (Cer et al., 2017). However, the STS systems performed much worse on the MT data than when tested on the Stanford Natural Language Inference (SNLI) Corpus data for the same language pair, with the top system achieving a Pearson correlation of only 0.34 compared to 0.83. These models are also complex, and for use in combination with fluency, we prefer a simpler approach for this study.

MT quality metrics are judged based on their correlation with human judgments, and recently, that has meant human adequacy judgments (Bojar et al., 2017).

BLEU (Papineni et al., 2002) is a widely accepted baseline measure of MT quality at the system level and, as such, is an obvious choice for a baseline adequacy metric. However, it may not be well suited for this task. Segments with high BLEU scores more closely match the reference, indicating high adequacy, but translations that receive a lower BLEU score may be inadequate, or they may be adequate with different word choice. For the purpose of detecting fluently inadequate translations, we can be confident that a segment with a high BLEU score is adequate, but low BLEU scores do not necessarily imply low adequacy.

To account for cases where a translation may be adequate but receive a low BLEU score, we need an adequacy metric less affected by word choice. This suggests the need to compare semantic representations rather than match strings. For our baseline vector-based metric, we use the common, simple approach of comparing sentence embeddings generated by averaging the word embeddings for each word in the sentence. However, this approach compares only the sum of the word vectors and does not directly compare them, so many unrelated sentences could produce the same sentence vector. We introduce an alternative word embedding-based sentence similarity measure that overcomes this flaw to produce a more reliable adequacy metric, bag-of-vectors sentence similarity (BVSS).

BVSS is an application of Saga (Similarity AGregation Application) introduced by Knox (2015). The Saga approach frames a task as an information similarity problem. Given a measure of information I for a multiset, the similarity between two multisets, X and Y, is the proportion of information from the union of X and Y that is found in both X and Y:

$$S(X,Y) = \frac{I(X) + I(Y) - I(X \cup Y)}{I(X \cup Y)} \tag{3.7}$$

The information measure in Saga uses single-linkage agglomerative clustering (Florek et al., 1951). If the items in a multiset are clustered according to similarity, more clusters indicate more disparate items and, therefore, more information. When we compare two multisets of items, X and Y, we first cluster each multiset separately to get $I(X)$ and $I(Y)$. We then pool all the items and cluster again to get $I(X \cup Y)$. If the items in X are similar to the items in Y, those items will cluster together, yielding fewer clusters than if they were different.

A nice feature of this approach is that in addition to the undirected similarity, we can modify Equation 3.7 to a directed form. The directed similarity to X of Y would be given by the proportion of information in Y that also appears in X:

$$S_X(Y) = \frac{I(X) + I(Y) - I(X \cup Y)}{I(Y)} \tag{3.8}$$

To compare sentences with this approach, we treat a sentence as a multiset of words and determine the similarity of words using the cosine similarity of their embeddings. Replacing X and Y in equation 3.7 with S for MT system output and R for reference gives us the $BVSS$ metric:

$$BVSS(S,R) = \frac{I(S) + I(R) - I(S \cup R)}{I(S \cup R)} \tag{3.9}$$

The directed form provides a way to measure when information is lost (i.e.,

the reference has more information than the MT output) or hallucinated (the MT output has more information than the reference). We will use *BVSS-reference* and *BVSS-system* to refer to these directed similarities. *BVSS-reference* is the proportion of the information in the reference that is also in the MT output, and *BVSS-system* is the proportion of the information in the MT output that is also in the reference:

$$BVSS_{reference} = \frac{I(S) + I(R) - I(S \cup R)}{I(R)} \tag{3.10}$$

$$BVSS_{system} = \frac{I(R) + I(S) - I(R \cup S)}{I(S)} \tag{3.11}$$

3.2.4 Detection Method Evaluation

Since there is no existing dataset with manual annotation of fluently inadequate translations, we first evaluate our fluency and adequacy prediction approaches comparing against direct assessment scores from WMT16 (Bojar et al., 2016) as 2016 was the only year in which human fluency judgments were collected. We then use our automated fluency scores on reference translations and automated adequacy scores on synthetic low adequacy "translations" to determine thresholds for high fluency and dubious adequacy.

3.2.4.1 Fluency Experiments

For WMT16, fluency judgments were collected for Czech-English (CS-EN), German-English (DE-EN), Finnish-English (FI-EN), Romanian-English (RO-EN), Russian-English (RU-EN), and Turkish-English (TR-EN) in the News shared task.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	All
MeanLP	0.326	0.213	0.277	0.258	0.228	0.324	0.270
NormLP	0.413	0.267	0.253	0.228	0.274	0.250	0.280
$WordLPmin_1$	0.045	0.011	0.058	0.024	0.053	0.040	0.037
$WordLPmin_2$	0.288	0.230	0.214	0.212	0.244	0.207	0.231
$WordLP_{25\%}$	0.400	0.260	0.239	0.205	0.289	0.216	0.267
$WordLP_{50\%}$	0.322	0.269	0.226	0.197	0.257	0.208	0.244
$WordLP_{mean}$	0.382	0.294	0.267	0.227	0.300	0.257	0.286
$WordLP_{mid}$	**0.425**	**0.343**	**0.349**	**0.313**	**0.345**	**0.386**	**0.359**

Table 3.1: Pearson correlation between each of the fluency prediction metrics and the human fluency direct assessment scores for each language and across all languages.

Annotations were collected with the goal of system-level reliability, so many segments only have one judgment. To improve reliability, we use only segments with two or more judgments. Predicted fluency scores are based on a 5-gram language model. We built a 5-gram KenLM (Heafield, 2011; Heafield et al., 2013) language model using the monolingual news training data from WMT16.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	All
%Fluent	59.22%	59.70%	56.79%	58.04%	60.80%	48.81%	57.21%
Precision	65.35	63.36	59.62	62.06	66.22	52.37	61.42
Recall	90.97	87.29	91.56	92.20	87.67	87.77	89.38
F1	76.06	73.42	72.21	74.18	75.45	65.60	72.81

Table 3.2: Precision, recall, and F1 on fluent translations for $WordLP_{mid}$ on system outputs for each language pair and on all system outputs. The percentage of outputs labeled fluent based on the human fluency judgments is also provided for reference.

For each of the metrics described in section 3.2.2, we calculated the Pearson correlation with the direct assessment scores for each language pair data set and for all the data combined. Results are shown in Table 3.1. Although these correlations are lower than we would like, we find that for all language pairs and for the combined

data, $WordLP_{mid}$ yields the highest correlation, so we will use this formula for our fluency prediction metric.

Because our goal is to correctly label sentences as fluently inadequate rather than to provide an exact score, we must select a fluency threshold τ_f to label a translation as "fluent". To determine this threshold, we computed the $WordLP_{mid}$ scores for the reference translation sentences in the WMT16 news training data. To cover most examples while allowing for variance in human judgments, the threshold is set at the point where 90% of reference segments would be labeled as fluent. Precision, recall, and F1 scores for $WordLP_{mid}$, are shown in Table 3.2. Across all data sets, we see high recall, but the precision is not as high. Although this suggests that this metric might overestimate the fluency of translations, we are more concerned with comparing between systems than with the raw scores.

3.2.4.2 Adequacy Experiments

We assess adequacy metrics using the direct assessment adequacy scores and system outputs for all language pairs from WMT16 (Bojar et al., 2016). Adequacy judgments were collected for all submitted systems in all language pairs in the News shared task. These annotations were used to determine the system rankings in the news task and as the gold standard quality judgments for the metrics shared task. For the metrics task, enough annotations were collected for each system-produced segment to establish segment-level reliability, while only enough judgments for system-level reliability were collected for the remainder of the segments for the

news task. Because we need segment-level reliability, we use only the metrics subset of the data as gold standard human judgments, and we use the reference translations from the news subset in generating synthetic inadequate examples.

We use the standardized human direct assessment adequacy scores from WMT16 (Bojar et al., 2016) as gold standard assessments in determining how well each adequacy metric correlates with human judgments. However, for binary questionable/acceptable adequacy judgments, we must be sure that the inadequate examples are clearly inadequate regardless of fluency and other MT quirks. The high correlation between human judgments of fluency and adequacy in Callison-Burch et al. (2007) and Graham et al. (2017) may indicate that human adequacy judgments are influenced by fluency, lowering the adequacy scores of disfluent translations. To ensure that our inadequate examples are truly inadequate, we rely on synthetic examples. We generate synthetic low adequacy translations by randomly selecting pairs of reference translations from the WMT16 news task and treating one as synthetic MT output and the other as the reference. We split these synthetic examples into dev and test sets. The dev synthetic examples are used in choosing the binary acceptable/questionable adequacy threshold τ_a as described below. The test synthetic examples are used as the questionable adequacy items in our adequacy precision/recall test set, with acceptable adequacy items chosen from actual WMT16 submissions. Because we are looking for extreme inadequacy and the systems in WMT16 were of competitively high quality, we use segments with direct assessment scores in the top 90% as acceptable adequacy in the test set.

For each metric defined in Section 3.2.3, we calculated the Pearson correlation

with the direct assessment scores for each of the WMT16 language pair data sets and for all the data sets combined (Table 3.3). Our vector-based metrics are based on word embeddings. We use the pre-trained aligned Wikipedia fastText word vectors (Joulin et al., 2018; Bojanowski et al., 2017). The averaged sentence embeddings had the lowest correlation across all language pairs. BVSS-System performed similarly well compared to BLEU, but BVSS and BVSS-Reference both outperformed BLEU.

	cs-en	**de-en**	**fi-en**	**ro-en**	**ru-en**	**tr-en**	**All**
BLEU	0.543	0.420	0.415	0.484	0.451	0.503	0.462
Avg Embeddings	0.439	0.190	0.312	0.363	0.235	0.303	0.296
BVSS	0.613	0.471	**0.519**	**0.562**	**0.555**	**0.589**	**0.543**
BVSS-Reference	**0.622**	**0.479**	0.490	0.556	0.519	0.563	0.536
BVSS-System	0.538	0.389	0.452	0.477	0.503	0.537	0.469

Table 3.3: Pearson correlation between each of the adequacy prediction metrics and the human adequacy direct assessment scores for each language and across all languages.

	Prec.	**Recall**	**F1**
BLEU	94.33	99.08	96.65
Avg Embeddings	84.56	**99.15**	91.28
BVSS	**99.39**	99.04	**99.22**
BVSS-Reference	99.00	99.03	99.01
BVSS-System	99.17	99.03	99.10
BLEU+BVSS	**99.61**	**99.81**	**99.71**

Table 3.4: Precision, recall, and F1 on BLEU, BVSS, BVSS-Reference, BVSS-System, and BLEU with BVSS and BVSS-System on the questionable adequacy test set with thresholds calculated based on predicted adequacy scores for the synthetic low adequacy dev data.

As with fluency, our goal for the adequacy metric is to correctly label a sentence as questionable adequacy rather than to provide an exact score. We used each candidate adequacy metric described in section 3.2.3 to score the segments in the

synthetic low adequacy dev set and set adequacy threshold τ_a for each metric such that 99% of dev set examples would be labeled inadequate. The precision, recall, and F1 on the synthetic test set using this threshold for each metric are shown in Table 3.4. We see that, as with correlation scores, the Averaged Embeddings have much lower precision than BLEU or any of the BVSS metrics, and the BVSS metrics have higher precision than BLEU.

Because of the potentially complementary differences in BLEU and BVSS, we also tested combinations of BLEU and the highest-performing vector-based metric, BVSS. We combine the metrics by marking a translation as questionable adequacy only if both metrics would label it as questionable. We see a slight improvement in F1 with the combination, and we adopt this metric for labeling segments as questionable adequacy.

3.2.5 System-Level Analysis of Fluently Inadequate Translations

Based on the fluency and adequacy evaluations in Section 3.2.4, we select $WordLP_{mid}$ and the BLEU+BVSS combination to label segment translations as fluently inadequate.

The results for segment-level fluency and adequacy prediction tests show that neither metric is perfect at the segment level. However, the impact of segment-level errors is lessened when segment-level scores are aggregated to compare across systems.

Koehn and Knowles (2017) showed that in out-of-domain and low-resource

	Arabic	**Chinese**	**Farsi**	**German**	**Korean**	**Russian**
Subtitles	30M	11M	6.2M	22M	1.4M	26M
UN v1	18M	-	-	-	-	-
WMT17	-	25M	-	5.8M	-	25M
LDC	1.3M	-	-	-	-	-
All General	49M	36M	6.2M	28M	1.4M	51M
TED	174K	169K	114K	152K	164K	180K
TED Test	1982	1982	1982	1982	1982	1982

Table 3.5: Number of segments in General Domain and TED training and test data for all languages

settings, NMT produces lower quality output than SMT, and they include examples where the NMT produced translations that were fluent but unrelated to the input. We seek to quantify this observation by estimating how often such fluently inadequate translations occur in SMT and NMT systems in different domain mismatch and training data settings. We score the output of 36 MT systems according to the percentage of fluently inadequate translations using the method described above.

We use a set of neural and phrase-based statistical MT models built from the same general domain data and adapted to translate a more specific domain, namely, transcripts of TED talks. We selected six languages to cover a range of resource availability scenarios and language families: Arabic, Chinese, Farsi, German, Korean, and Russian.

The number of segments of training and test data for each language is summarized in Table 3.5. The same tokenization was performed for all systems for a given language, and the tokenized data was split into subwords for NMT training using byte pair encoding (BPE) (Sennrich et al., 2016). The BPE models were trained separately on the source and target language with 30K BPE symbols.

	Arabic	German	Farsi	Korean	Russian	Chinese
Joshua General	23.50	30.65	13.41	6.34	24.49	14.79
Joshua TED Only	24.49	28.72	16.56	9.81	21.85	13.32
Joshua Adapted	27.11	31.35	17.71	10.24	25.23	15.70
Sockeye General	29.6	34.59	22.22	11.56	**28.6**	15.92
Sockeye TED Only	27.42	32.25	21.31	14.4	22.9	16.18
Sockeye Adapted	**35.37**	**39.9**	**27.92**	**17.22**	**28.6**	**20.37**

Table 3.6: BLEU scores for all systems

All languages used data from the OpenSubtitles[1] corpus (Tiedemann, 2009) in the General domain training and dev data sets. The Chinese, German, and Russian models used additional parallel corpora from WMT17[2] (Bojar et al., 2017). For the Arabic models, we added data from the Linguistic Data Consortium (LDC)[3] and the UN v1 corpus[4] (Ziemski et al., 2016).

The domain for the In-Domain and Domain-Adapted models was TED talks. Training, dev, and test sets for the domain were from the Multi-target TED Talks Task (MTTT) corpus (Duh, 2018). All systems, regardless of training setting, were tested on the TED domain test set.

Fluency scores for each system were generated based on a language model built on the English side of its primary training data. As noted in Section 3.2.2, it is important that the language model match the training data, and we expect this to be particularly true when the test set is out-of-domain. We, therefore, use only the General domain data for both the General models and the adapted models, while the In-Domain models use the in-domain training data. Thresholds were calculated

[1] `https://www.opensubtitles.org/`
[2] `https://www.statmt.org/wmt17/translation-task.html`
[3] LDC2004T18, LDC2007T08, and LDC2012T09
[4] UN v1 is included in the Russian and Chinese WMT17 data

similarly to the thresholds on the WMT16 data: thresholds for $WordLP_{mid}$ were calculated based on the General domain training data and thresholds for sentence BLEU and BVSS based on synthetic data built from the TED training data.

The statistical systems were built using the Apache Joshua toolkit[5] (Post et al., 2015). We tested three SMT models for each language: Joshua General, Joshua In-Domain, and Joshua Domain-Adapted, which were trained respectively on the General domain data, on the TED training data, and on both. Language models for all systems were built from the English side of the training data. The Domain-Adapted model was tuned on TED dev data.

The neural systems were built using Sockeye[6] (Hieber et al., 2017). The systems used two LSTM layers in both encoder and decoder with hidden size 512 and word embeddings dimension 512. We used a batch size of 4096 and created a checkpoint every 4000 mini-batches. Our systems employed the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0003. As with the SMT, we built three models for each language: Sockeye General, Sockeye In-Domain, and Sockeye Domain-Adapted. The Sockeye General and In-Domain models were trained with the same data as the corresponding SMT models. The Sockeye Domain-Adapted models were trained using continued training on TED data starting from the Sockeye General model as in Luong et al. (2015) and Freitag and Al-Onaizan (2016).

[5] `https://cwiki.apache.org/confluence/display/JOSHUA/`
[6] `https://github.com/awslabs/sockeye`

3.2.6 System Analyses

We compute the percentage of fluently inadequate translations in the system output of all MT and SMT systems to determine the effect of MT paradigm and training data on the occurrence of fluently inadequate translations.

Although the language pair and system vary, we can directly compare the output of the systems because the test data for all systems is from the *Multi-target* TED corpus. Note that in the corpus, the source is English and the other languages are translations while our task is translating into English. This means that if there are human translation errors or non-literal translations, the source will be inconsistent across languages, but the reference will be the same. Table 3.7 shows English references for two different segments in Farsi and Chinese that yielded fluently inadequate MT output, along with their corresponding source and system outputs. For some segments, the human translation (our source) may have a slightly different meaning from the original (our reference), but the fluently inadequate examples we seek to identify are much further in meaning from both the source and reference. For instance, in the Chinese-English example, the Chinese adds information that must be inferred from context in the original English. The Chinese text literally translates to "Crow parents also teach their children these kinds of skills." The Sockeye TED and Joshua outputs reflect this additional information, but the Sockeye General output is fluent but completely unrelated to the reference.

Figure 3.1 shows the percentage of segments labeled as fluently inadequate for each system. Even the highest percentage (Chinese-English Sockeye General)

System	FA-EN Example	ZH-EN Example
Source	انگيزه هاي زيركانه تري داشته باشيد	乌鸦父母还教会自己的孩子这样的技巧呢。
Reference	get smarter incentives.	parents seem to be teaching their young.
Joshua General	terry زيركانه have motives.	parents also teach their children the skills like this?
Joshua TED	you have the انگيزه زيركانه needed	parents crow can also teach our kids that the skills that.
Joshua Adapted	terry motives inspired.	the parents teach their children such skills.
Sockeye General	have a more subtle motivator.	**i 'm afraid i 'm not going to have to go to bed.**
Sockeye TED	**there's a lot of gamers.**	and the crow parents taught their kids like this.
Sockeye Adapted	have smarter motivations.	and their parents also taught their children how to do it.

Table 3.7: Reference translation and example translations from the Farsi-English and Chinese-English systems. Fluently inadequate examples are bold.

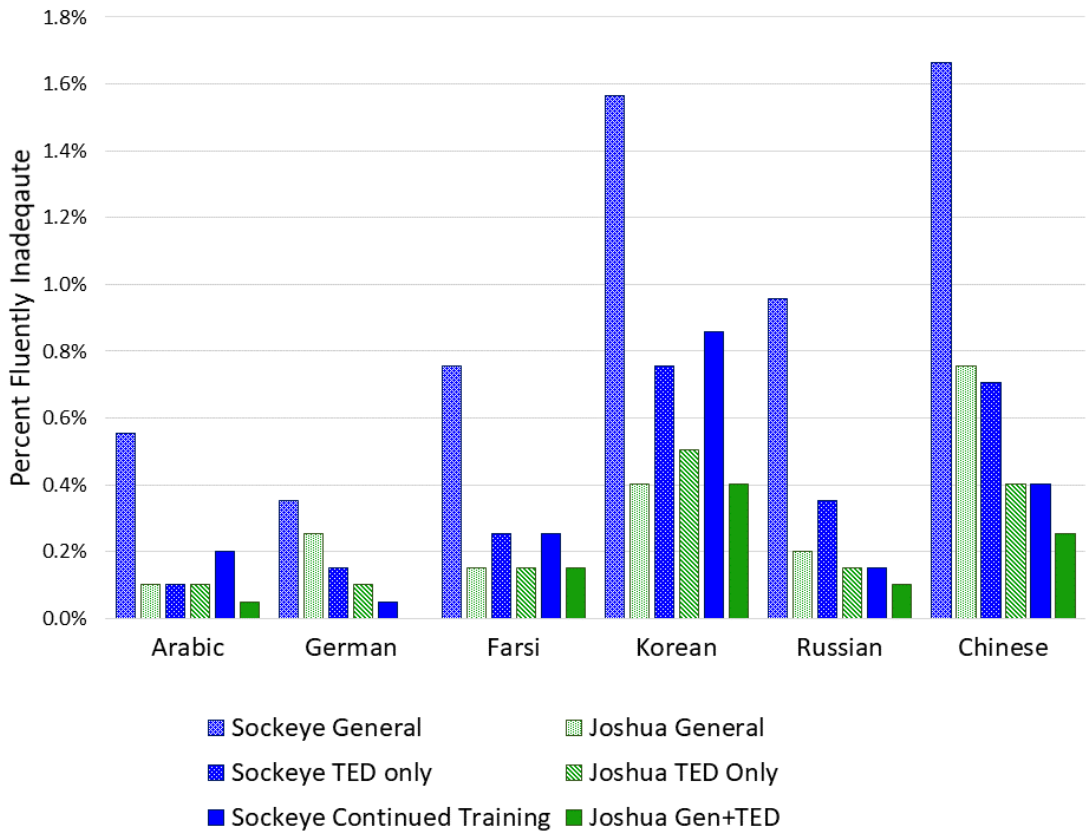


Figure 3.1: Segments labeled as fluently inadequate for General, TED, and Domain-Adapted Sockeye and Joshua models for all languages.

is less than 2%. Based on the high recall and low precision scores for the fluency metric in Section 3.2.4.1, we expect that we are overpredicting fluently inadequate translations, so the actual percentage may be even lower. This confirms that these errors are indeed rare.

We also see from Figure 3.1 that the NMT models for Korean and Chinese, the languages most typologically different from English, have the highest levels of fluently inadequate translations on out-of-domain models. Although they have similarly high percentages of fluently misleading and similar amounts of in-domain training data, the Chinese domain-adapted model improves much more than the

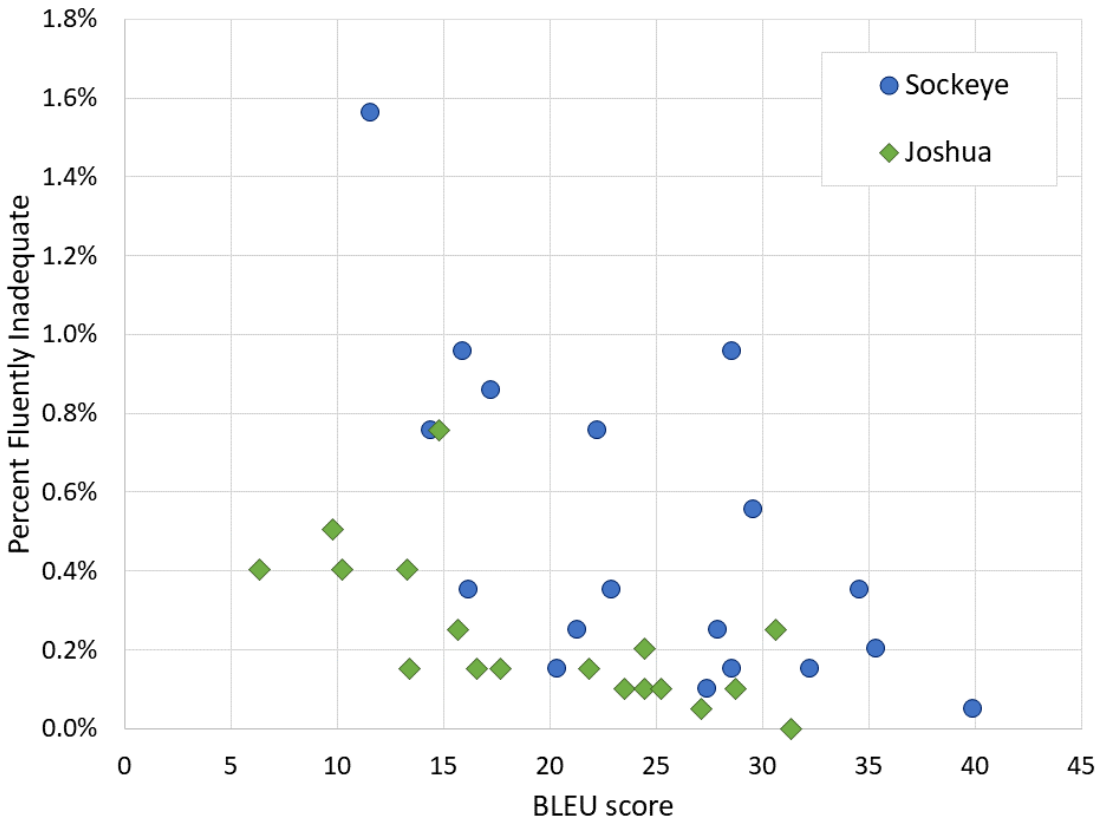


Figure 3.2: Segments labeled as fluently inadequate vs. BLEU score for all Sockeye and Joshua models for all languages.

Korean domain-adapted model.

We compare the percent fluently inadequate segments to system BLEU scores in Figure 3.2. Based on the definition of our metric for fluently inadequate translations, translations with high sentence BLEU cannot be labeled fluently inadequate, so we expect a strong negative correlation between system BLEU and the percent fluently inadequate. We do see this negative correlation, but we can also see a clear difference in the percent fluently inadequate for the SMT vs NMT systems.

The NMT systems with low BLEU scores have much higher percentages of fluently inadequate translations than the similarly low-scoring SMT systems. This

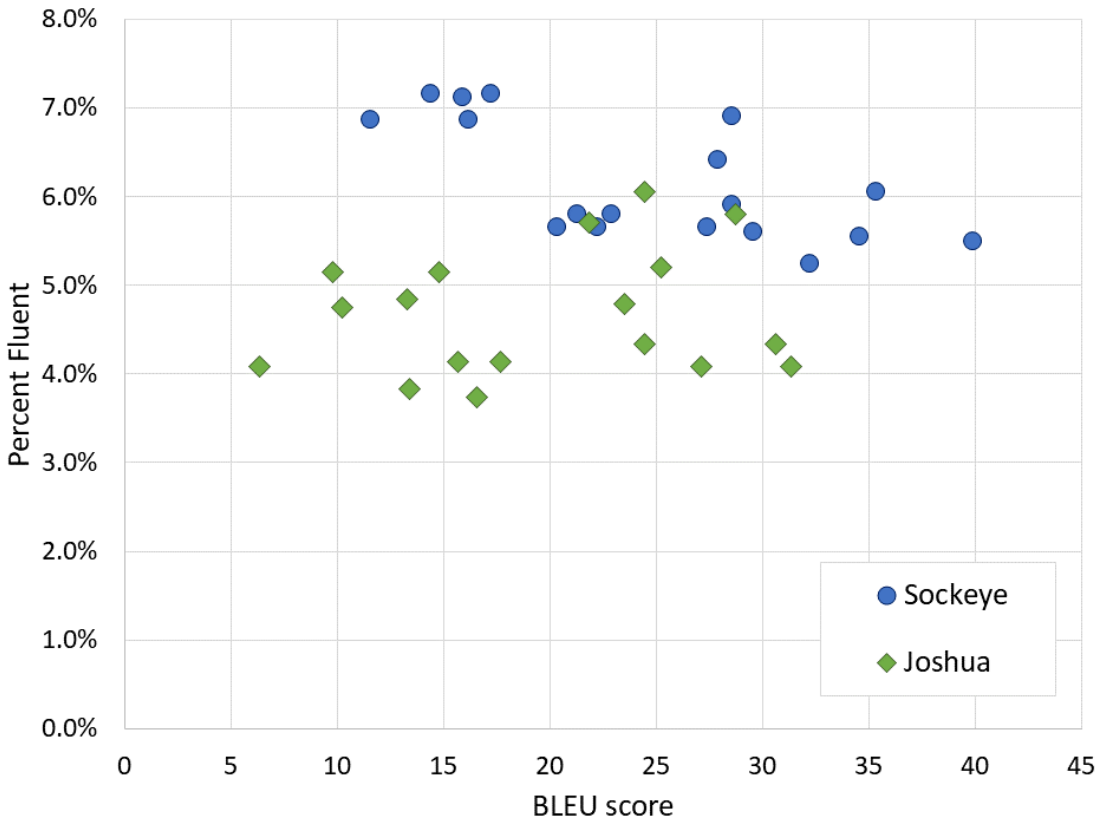


Figure 3.3: Segments labeled as fluent vs. BLEU score for all Sockeye and Joshua models for all languages.

follows the suggestion in Koehn and Knowles (2017) that NMT is more prone to producing output that is disconnected from the source text when trained with insufficient or out-of-domain data. Indeed, we can see in Figure 3.1 that the NMT consistently has a higher percentage of fluently inadequate translations than the SMT.

Because our fluency metric relies on language models very similar to the language models used in the SMT systems, we might suspect that the fluency metric is biased towards the SMT models, potentially making SMT output more likely to be labeled as fluently inadequate. However, Figure 3.3 shows that the NMT systems

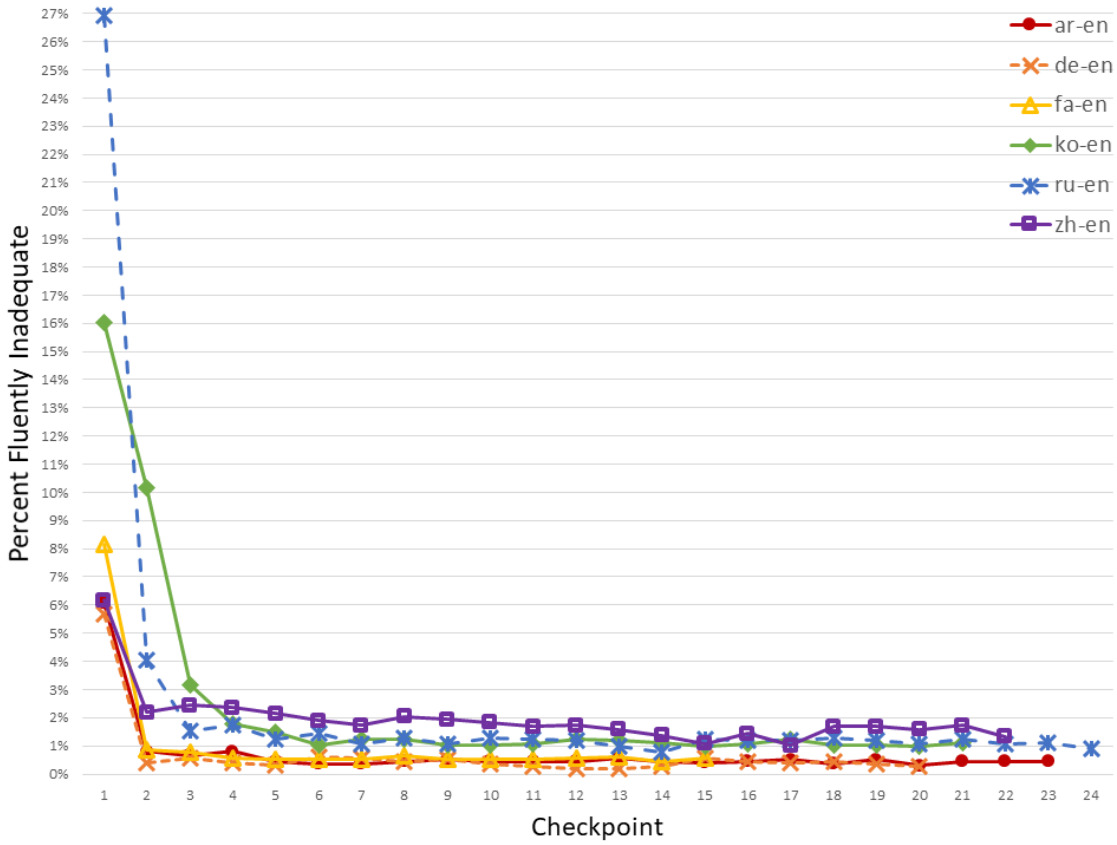


Figure 3.4: Percent fluently inadequate at each checkpoint during in-domain training

still consistently have more segments labeled as fluent compared to SMT systems with similar BLEU scores. This agrees with prior work showing that NMT output is more fluent than SMT and suggests that while the fluency metric likely leads to overprediction of fluently inadequate translations, it does not do so in a way that favors one paradigm over the other.

We also measured the percentage of fluently inadequate translations on the development set during training. Figure 3.4 shows that the percent fluently inadequate levels off very quickly, flattening after a few checkpoints on the in-domain model.

3.2.7 Conclusion

We have introduced an approach to automatically detect fluently inadequate translations in machine translation output based on automatic fluency and adequacy metrics. Applying this technique to a diverse set of statistical and neural MT systems, we found that although fluently inadequate translations are rare, NMT does appear to be consistently more prone to this type of error compared to SMT. Improving the match between train and test conditions with continued training on in-domain data reduces these errors. These findings raise a key question: How often are fluently inadequate translations actually misleading to human users? The next chapter introduces an experiment to test this question and determine whether fluency alone is sufficient for predicting whether users will believe that the output of an MT system is likely to be accurate.

Chapter 4: Understanding MT Misleadingness: Believability

The previous chapter looked at potentially misleading MT output from the perspective of traditional intrinsic MT evaluation features, fluency and adequacy. The pilot study suggested that fluently inadequate translations may mislead users, and the experiment of Chapter 3 indicated that fluently inadequate translations are more common in neural MT than in previous technologies. However, both assume that fluency is sufficient to cause an inadequate translation to be misleading. In this chapter, we take a user-centered view of MT evaluation, exploring one aspect of users' perception of MT: *believability* of the output, defined as a monolingual user's perception of the likelihood that the meaning of the MT output matches the meaning of the input, without understanding the source.

4.1 Introduction

Assessing the degree to which MT is believable acknowledges that users play an active role in interpreting MT output, informed by their linguistic competence, common-sense reasoning abilities, and knowledge of the world. What we learn from assessing believability can complement traditional evaluation methods to inform the deployment and even development of MT systems, particularly for gisting and

communication use cases.

We first define believability in the context of MT and relate it to concepts from the credibility literature, such as plausibility, and concepts from the MT evaluation literature, such as fluency. We apply MT direct assessment (DA) protocols to obtain human judgments of believability, annotating the output of neural machine translation (NMT) systems for three challenging language pairs (Arabic-, Farsi-, and Korean-to-English) with varying translation quality. These annotations show that believability is closely related to, but distinct from, fluency. Finally, we conduct a qualitative analysis that suggests that in addition to fluency features, believability is also influenced by semantic features.

We define MT *believability* as a user's perception of the likelihood that the meaning of a given MT output matches the meaning of the input, *without understanding the input*. Whether the user accepts the output unquestioningly or finds it unbelievable, their judgment will affect how they act on it, regardless of the true accuracy of the translation. For example, a Facebook user might be dubious of a translation from Chinese with the phrase "blowing a little more cow" and ask the author for clarification and learn that it is a literal translation of an idiom, "吹牛" (meaning to brag). Users may take more consequential action, such as the Israeli police officers who chose not to consult an Arabic speaker before arresting a man based on a believable mistranslation of his "good morning" Facebook post as "attack them" (Berger, 2017).

To illustrate how believability can be independent of adequacy, Table 4.1 shows examples of different levels of believability for translations at different levels of ad-

	Less Adequate	More Adequate
More Believable	"putnik" [sic] was an interactive film.	"Sputnik" was in the city center, the negative. It was not affected.
Less Believable	spaghetti, the nigerian was in the middle of the city, he didn't touch it.	sputnik was downtown, didn't look, never touched.

Table 4.1: Machine Translations of human translations of a line from a TED talk discussing the loss of film negatives in a fire (Hoffman, 2008). One film "Sputnik" was spared from the fire because it was not in the building. Original text: *"Sputnik" was downtown, the negative. It wasn't touched.*

equacy based on our annotations. The translations on the right (More Adequate) convey key information: a named entity (*sputnik*), its location (*downtown*), and the fact that it was not affected, but a user might not accept the information in the bottom-right translation because it is not believable. The less adequate translations (left) are missing important information and also include incorrect information. The bottom-left translation is not believable, so a monolingual user would not be misled by it. However, the more believable top-left translation might mislead a monolingual user.

Because these judgments are based on perception, they may be more subjective than traditional MT DA features. We control for some factors that may affect believability (Section 4.3), resulting in annotations that are similarly reliable to the DA features (Section 4.4).

4.2 Defining Believability

Although believability is an unexplored aspect of MT, there is prior work outside of MT on the general concept of *credibility*. As discussed in Section 2.1, common elements of credibility include source, media, and message credibility (Rieh and Danielson, 2007). For MT, we can think of the MT provider as the source, the interface through which the MT is viewed as the medium, and the output itself as the message or content. All of these aspects likely affect the credibility of deployed MT systems, but our investigation of MT believability is focused on the content (MT output). Some intrinsic content features addressed in the credibility literature that may affect MT believability include reasonableness (Liu, 2004; Kim and Oh, 2009; Kim, 2010; John et al., 2011) and grammatical errors (Fogg et al., 2001; Everard and Galleta, 2005; Metzger et al., 2010; Chesney and Su, 2010; John et al., 2011).

Reasonableness is a semantic feature encompassing elements of plausibility, logic, and internal consistency. These elements are related to previously studied concepts in the Computational Linguistics literature: semantic plausibility, commonsense reasoning, and discourse coherence. Semantic plausibility can be thought of as "whether in an ordinary real-life situation (not "fairy-tale" circumstances) the sentence could be reasonably uttered" (Kruszewski et al., 2016). If the source text is expected to reflect "ordinary real-life," the output should be plausible to be believable. MT output may also be unbelievable if it violates commonsense reasoning, a challenging element of Natural Language Understanding (Mostafazadeh et al., 2016). Lack of discourse coherence might likewise signal unbelievable translations.

For document generation, improving the consistency of generated documents makes it harder for human subjects to distinguish automatically generated text from real text (Karuna et al., 2018).

Grammatical errors are related to *fluency*, a traditional MT quality evaluation feature. Fluency has been defined as a judgment of "whether the translation reads like good English...without knowing the accuracy of the content," and is typically combined with *adequacy*, an assessment of "the degree to which the information in a professional translation can be found in an MT (or control) output of the same text" (White et al., 1994). A user who cannot understand the source cannot judge adequacy but may use expectations based on features like fluency and reasonableness to guess. Believability could thus be seen as *predicted adequacy* via a *human cognitive process* with inputs from surface features of the output, such as fluency and semantic cues from context.

Two other Computational Linguistics concepts that relate to both reasonableness and grammatical errors may affect believability: acceptability and comprehensibility. In empirical linguistics, acceptability judgments measure users' linguistic competence (Schütze, 2016). While acceptability is primarily used to observe grammatical knowledge, judgments are not limited to grammaticality in practice: "semantic plausibility, various types of processing difficulties, and so on, can individually or jointly cause grammaticality and acceptability to come apart" (Lau et al., 2017). These breakdowns can lead to issues with comprehensibility. Popović (2020b) cites comprehensibility as a key factor in the misleadingness of MT output: if a user cannot understand the text, they cannot be misled by it. Similarly, if they

cannot understand it, the user is unlikely to believe that the translation is correct.

4.3 Annotating Believability

To understand the relationships between believability and traditional MT quality criteria (fluency and adequacy), we hired professionals to annotate a selection of MT outputs from the previous chapter in tasks based on the fluency and adequacy DA methods of Graham et al. (2013) and Bojar et al. (2016). The annotated data sets are available at: `https://github.com/mjmartindale/mt_believability` .

Our annotators were salaried translators with proficiency levels of at least Advanced on the ACTFL scale (ACTFL, 2012) rather than MT researchers or crowd workers as in WMT (Barrault et al., 2020). Because they were not paid per item, they were willing to spend significant time on each item, averaging 15 items in 30 minutes compared to WMT annotators who averaged 100 items in that time (Bojar et al., 2018). We believe this reflects more attention to detail as indicated by the correlation between annotators (see Section 4.4). We note that factors such as foreign language proficiency may affect believability judgments. Further work with a wider variety of annotators is needed to identify and quantify those effects.

We followed segment-level DA scoring best practices established by WMT (Bojar et al., 2016, 2017, 2018). For all three features (believability, fluency, and adequacy), we used 100-point sliding scales to correct for annotator bias by normalizing scores (Graham et al., 2013). Each question asks, "How much do you agree with the following statement?" then provides a statement about fluency, adequacy,

or believability. The fluency and adequacy questions were taken directly from the only WMT evaluation that included fluency direct assessment, WMT16 (Bojar et al., 2016). The believability question uses our definition of believability with an introductory phrase to assure the annotator that we understand that it is not possible to truly evaluate the meaning without the source: "Even without having seen the source text, I believe the *meaning* of this translation is likely to match the *meaning* of the original." The phrase "Even without having seen the source text" was included in the prompt in response to feedback from translators who reviewed the task design and said that annotators would protest that they could not judge the meaning of the translation without seeing the source text. Annotations were performed using the Turkle[1] annotation platform. Figures 4.1 and 4.2 provide screenshots of the annotation interface.

Long documents were broken up into salient chunks, and segments were annotated in their original order to provide discourse context, as in WMT19 "Segment Rating + Document Context" (Barrault et al., 2019). Annotations were performed using a localized crowdsource platform similar to Amazon Mechanical Turk. For each chunk, annotators first scored fluency and believability based only on the MT output. They then scored the same segments for adequacy, given both source and output.

For each segment, we calculate a z-score and a label for each feature. We calculate scores following Bojar et al. (2018). Each annotator's raw scores are converted to z-scores based on their own mean and standard deviation, and the z-scores

[1] `https://github.com/hltcoe/turkle`

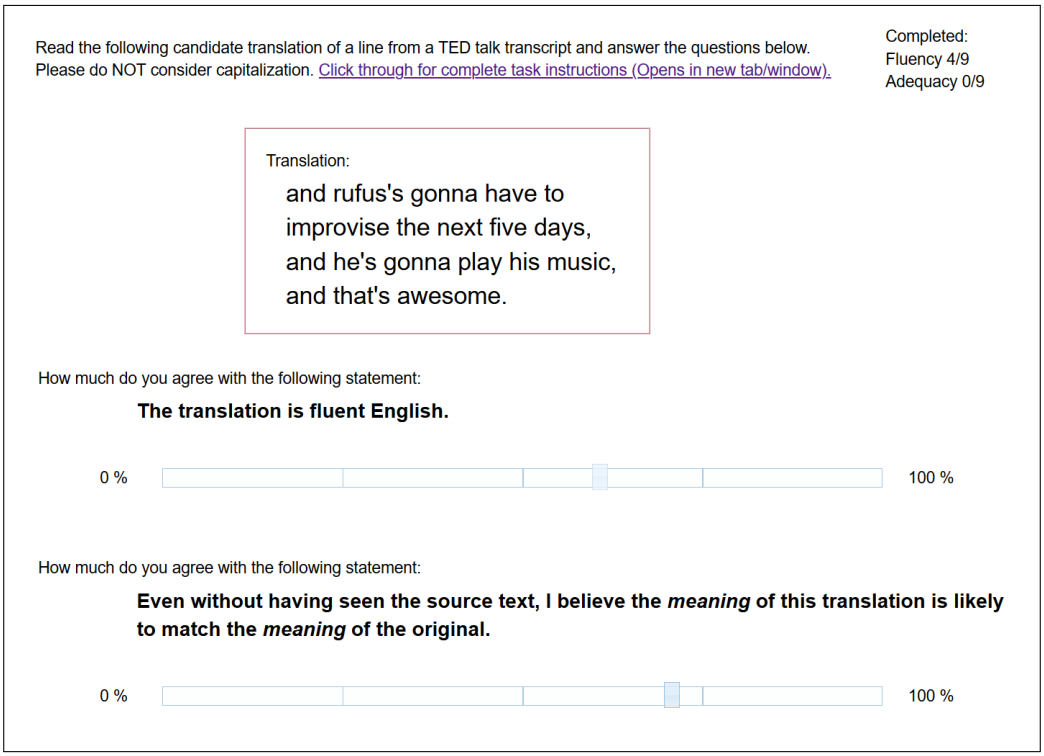


Figure 4.1: Example question from monolingual annotation phase, fluency and believability

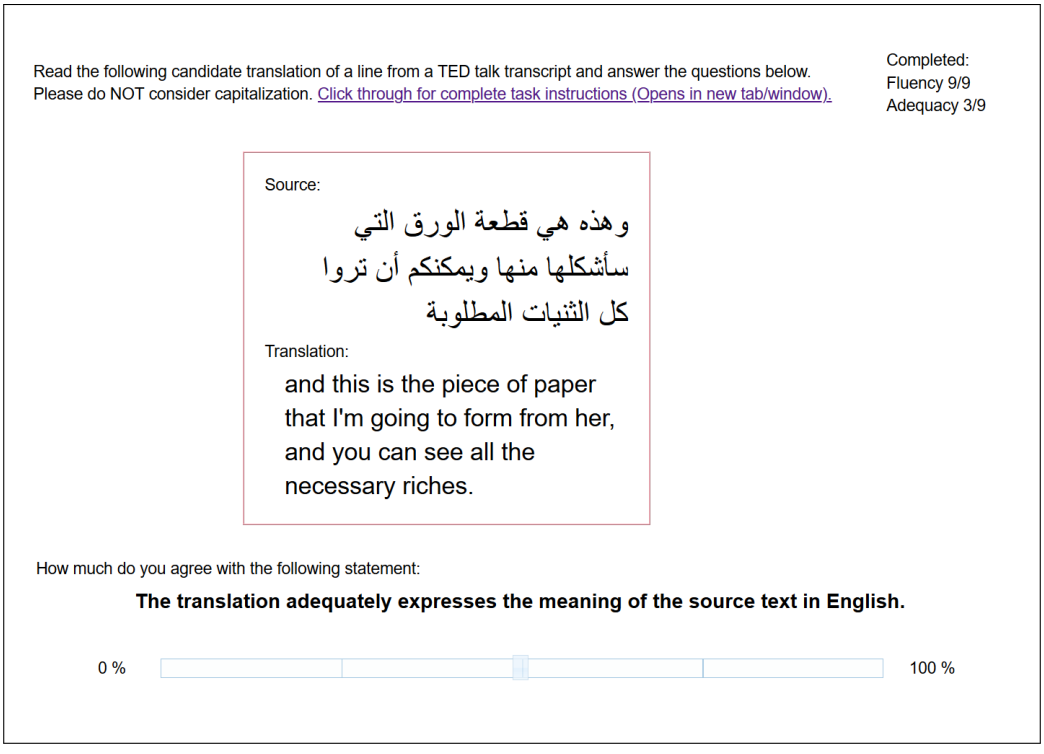


Figure 4.2: Example question from bilingual annotation phase with adequacy question

for each segment are averaged across annotators. Segments with positive z-scores are labeled TRUE, and negative z-scores are labeled FALSE.

We chose a test set that is comparable across three typologically different languages with different amounts of training data. Our test data comes from The Multi-Target TED Talks Task (MTTT)–a collection of bitexts across 20 languages, including both training and test partitions (Duh, 2018). The test set is fully sentence parallel across all languages, with original talk transcripts as the English and human translations for the other languages. Because our annotators are native English speakers, they are more qualified to judge the fluency of English output than text in their second language, so we use the non-English translations in MTTT as "source," and machine translate them into English. [2] In the test set, there are 29 talks totaling 1,982 segments, but we exclude one talk ("Nellie McKay sings 'Clonie'") that is too poetic for MT. The final set is 1,976 segments.

Because our goal is to examine *segments* annotated for believability, fluency, and adequacy judgments rather than to *compare* systems, we need MT that will produce outputs across a range of quality. Output that is inadequate but believable is of particular interest, so we rely on estimates of the distribution of "fluently inadequate" translations on MTTT from the work in Section 3.2 and use the same Arabic, Farsi, and Korean "general" models to capture the range of training data sizes and output quality we believe will provide interesting examples for our analysis. The training data is 49M, 6.2M, and 1.4M segments in Arabic, Farsi, and

[2] This may introduce "translationese" effects in the input text. While it might affect the distribution of error types, our focus is on the annotations themselves, not their distribution.

Korean, respectively. The systems are built in Sockeye (Hieber et al., 2017) using the 'SockeyeNMT rm1' settings from the MTTT leaderboard[3]. The resulting systems achieved BLEU (Papineni et al., 2002) scores of 26.6 for Arabic, 22.2 for Farsi, and 11.6 for Korean.

4.4 Quantitative Analysis

Tables 4.2 and 4.3 report the number of annotators for each segment and the number of segments annotated for each annotator. 63 translators participated in the annotation process. Korean-to-English and Arabic-to-English had the most annotators (26 and 27) and the highest number of annotators per segment (10 for Korean and at least 7 for Arabic). There were three annotators whose annotations were deemed unreliable due to low correlation with the mean (< 0.5). Only 10 annotators were available for Farsi, resulting in fewer annotators per segment (median: 4) but more segments per annotator (median: 802), making the z-score process more reliable. Although these are lower than the target of 15 annotators for reliable crowd-sourced segment-level annotations identified by Graham et al. (2015), our professional annotators are generally more reliable than crowd workers and, indeed, after excluding the questionable annotators, we see a strong correlation of individual annotator scores with the mean as shown in Table 4.4, along with Fleiss' Kappa scores for inter-annotator agreement for the thresholded labels of Fluent (or not), Believable (or not), and Adequate (or not). For each language, we see similar kappa scores and average correlation with the mean for fluency and believability.

[3] https://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/

This suggests that among our annotators, believability is no more subjective than fluency.

	Arabic	**Farsi**	**Korean**
Annotators	27	10	26
Min segs	9	47	28
Mean segments	822	760	
Median segs	525	802	400
Max segs	1925	1976	1976

Table 4.2: Segments annotated per annotator

Annotators	**Arabic**	**Farsi**	**Korean**
2+	1976	1976	1976
4+	1976	1439	1976
8+	1976	48	1976
10	1713	0	1976
Mean Annotations	3	4	10
Median	9	4	10
Max	10	8	10

Table 4.3: Annotations per segment

	Arabic			Farsi			Korean		
	FL	BL	AD	FL	BL	AD	FL	BL	AD
Mean Corr.	0.698	0.689	0.728	0.809	0.793	0.830	0.773	0.754	0.793
Std dev	0.137	0.122	0.129	0.087	0.099	0.083	0.112	0.120	0.076
κ	0.423	0.363	0.346	0.236	0.250	0.438	0.440	0.426	0.404

Table 4.4: Annotator correlation with the mean and Fleiss' Kappa scores for fluency (FL), believability (BL), and adequacy (AD)

Although the distribution of labels is specific to this set of output, it provides context for the other results. The first three rows of Table 4.5 show the percentage of segments with each label. We see that the percentage of positive examples for each label roughly relates to the system BLEU score, with Arabic having the highest

	%Arabic	%Farsi	%Korean	%All
Fluent	57.9	50.4	49.4	52.5
Believable	61.0	51.2	49.0	53.7
Adequate	59.4	52.4	44.1	52.0
Inadequate+Believable	25.9	19.4	21.8	22.2

Table 4.5: Percent of segments with each label (rows 1-3) and percent believable but inadequate (potentially misleading) segments (row 4).

and Korean having the lowest.

Table 4.6 shows the Pearson correlations between the scores for fluency, believability, and adequacy. The Believability:Adequacy relationship is important because inadequate believable translations may mislead monolingual users. The Believability:Adequacy correlation is higher than the Fluency:Adequacy correlation across all languages. This may be particular to the nature of TED talk transcripts, or it might reflect the influence of context. Adequate segments may fit the context well enough to be believable, even if not fluent. The same trend is reflected in the fourth row of Table 4.5, Inadequate+Believable. Most inadequate translations are not believable, but 19-25% are believable, making them potentially misleading. Arabic has the highest rate of these potentially misleading translations. The larger training data may improve both translation and generation, improving overall quality but enabling more believable errors. However, the lower quality Korean also has a higher percentage potentially misleading than Farsi. This could mean that all

	Arabic	Farsi	Korean	All
Fluency:Believability	0.89	0.96	0.97	0.94
Believability:Adequacy	0.71	0.74	0.75	0.73
Fuency:Adequacy	0.62	0.73	0.72	0.69

Table 4.6: Pearson correlation between fluency, believability, and adequacy.

results are idiosyncratic to this data, or perhaps the relationship is bimodal.

As expected, Fluency and Believability have the strongest correlation. This indicates that the two features are closely linked, but they are not identical, particularly in the case of Arabic. Table 4.7 illustrates this distinction. The top two rows show the percentage of fluent translations that are believable and unbelievable, while the bottom two rows show the percentage of disfluent translations that are (un)believable. Over 90% of segments labeled as fluent are also believable. For Farsi and Korean, disfluent segments are unbelievable over 90% of the time, but for Arabic, this happens for only 81.7% of the segments. If one were to attempt to identify potentially misleading translations using fluency as a proxy for believability, as in Section 3.2, most segments would be correctly labeled, but 6-8% of translations that are fluent but not believable would be incorrectly labeled as misleading, while nearly 20% of segments from the Arabic system that are believable but not fluent would be missed.

	%Arabic	%Farsi	%Korean	%All
Fluent believable	92.1	93.8	93.8	93.1
Fluent unbelievable	8.0	6.2	6.3	6.9
Disfluent believable	18.3	8.0	5.4	10.1
Disfluent unbelievable	81.7	92.1	94.6	89.9

Table 4.7: Percent of fluent and disfluent segments that are believable or unbelievable.

4.5 Qualitative Analysis

An informal examination of a small sample of unbelievable segments suggested that those labeled as unbelievable often fail semantically, with strange phrases (e.g., "kidney steel", "iron code"), illogical clauses (e.g., "the only natural thing is that my son is the vending machine"), or unlikely argument structure (e.g., "to prove that it seems impossible," "...a state of treason for a raven, which explains that he's cute"). Unbelievable translations may also be syntactically grammatical but semantically unintelligible (e.g., "if you want a long time, I'm actually doing something about it."). These observations suggest that, in addition to fluency, believability may be tied to comprehensibility and plausibility.

We validated the initial observations with an analysis of 400 segments consisting of approximately 100 segments per combination of believability and fluency from the original annotation task. We recruited and trained three annotators with backgrounds in linguistics and extensive experience working with MT output. Two annotators labeled all 400 segments, and the third annotator reviewed and reconciled the two labels when there was disagreement. Grounded in the initial examination of the small sample of segments, we identified issues related to comprehensibility, plausibility, and fluency with which to label the MT outputs, as well as labels for specific types of MT errors easily detected based on the MT output alone. Each of these error types is described in detail below.

The number of segments with each category of error label, how many of those segments were labeled as believable or fluent during the original annotation pro-

cess in Section 4.3, and the Cohen's kappa inter-annotator agreement score for each label before reconciliation is shown in Table 4.8. Kappa scores are highest (over 0.5) for the largely unambiguous MT-specific error categories (with the exception of Bad copy, of which there was only one example). Agreement on fluency (κ=0.441) is comparable to fluency agreement in Graham et al. (2013) and plausibility agreement (κ=0.284) is within the range reported in Lambert et al. (2010). The inter-annotator agreement for span-level intelligibility annotations in Popović (2020a) cannot be compared to our token-level annotations, but we expect our intelligibility agreement (κ=0.346) to be somewhere between plausibility and fluency due to the semantic and syntactic elements. Semantic intelligibility agreement (κ=0.0.269) is similar to plausibility agreement, but syntactic intelligibility agreement is, alarmingly, negative (κ=-0.239) although the simple agreement was 94% (23 disagreements in 400 examples). The much higher agreement across all intelligibility issues suggests that this is largely because of the low number of examples with syntactic intelligibility issues (8 out of 400) leading to a stricter kappa (Feinstein and Cicchetti, 1990; Delgado and Tibau, 2019), but the fact that those few disagreements lead to a negative kappa rather than a near-zero kappa may also reflect that significant syntactic errors may be more or less difficult for one person to interpret than another. In this case, although both annotators had similar educational backgrounds, they specialized in different language families (Sino-Tibetan and Austronesian) and may each have been more open to "word salad" translations with similar patterns to languages in their respective language families.

The initial comprehensibility labels were: *Understandable* ("This segment

has only one fairly easily recognizable interpretation, even if that interpretation is not a reasonable statement in a real-world context."), *Ambiguous* ("This segment has a couple fairly easily recognizable, specific interpretations."), and *Unintelligible* ("There are many possible meanings or it is very difficult or impossible to glean any meaning from the segment."). Segments labeled as Ambiguous or Unintelligible are grouped together under the category of **Intelligibility** issues. Segments with intelligibility issues were also labeled according to whether the source of the difficulty in understanding was **Semantic**, **Syntactic**, or both. Of the examples labeled as having intelligibility issues (51 or 12.75%), nearly all were labeled as having semantic difficulties (46 or 90.2%). Only 9 segments with intelligibility issues (17.65%) had been previously labeled as believable. This is consistent with the assertion in Popović (2020b) that unintelligible translations are unlikely to mislead MT users. These segments were also overwhelmingly considered disfluent (80.39%) in the original annotation, and 88.24% (45) of the segments with intelligibility issues were also labeled as having grammatical issues. Although intelligibility issues would still be covered by a fluency-based definition of believability, we will see that these issues are a stronger indicator of unbelievable translations than other kinds of fluency errors.

The plausibility labels were: *Plausible as written* ("This segment, as written, makes sense in context."), *I know what it means* ("The meaning as written is implausible, but a reasonable reader could guess the correct meaning."), *Implausible* ("This segment does not make sense in context."), and *n/a* ("This segment is too unintelligible to have any meaning, so plausibility cannot be judged."). Due to the inconsistency and subjectivity of the more specific plausibility issue labels, all

Issues	κ	Segments	Believable	Fluent
Intelligibility (all)	0.346	51	9	10
↳ Semantic	0.269	46	8	8
↳ Syntactic	-0.239	8	1	1
Plausibility	0.284	156	36	58
Fluency	0.441	208	92	42
MT-specific (all)	0.547	31	13	8
↳ Duplication	0.695	19	10	4
↳ Non-words	0.523	12	2	3
↳ Bad copy	-0.005	1	1	1

Table 4.8: Interannotator agreement, number of segments with each category of issue, and how many of those segments were labeled as believable and as fluent during the original annotation process.

labels other than *Plausible as written* are combined into a single category of segments with **Plausibility issues**. Based on the plausibility error types reported in Ko et al. (2019), annotators could also note whether plausibility issues arose from within-segment context or between-segment context and whether the content was contradictory (inconsistent with context) or non-sequitur (completely detached from context). These labels proved difficult to assign consistently and were not used in the final analysis. Overall, 39.00% of segments (156) were labeled as having plausibility issues. Most segments with plausibility issues were labeled as unbelievable (76.92% or 120). This matches our intuition that plausibility is tied to believability but leaves room for subjectivity in both plausibility and believability judgments.

The fluency labels were: *Grammatical error* ("The text includes one or more grammatical errors such as wrong grammatical category, verb disagreement, incorrect function word, wrong pronoun, missing or extra determiner, etc.") and *Awkward* ("This segment contains one or more words, phrases, or combinations of words or phrases that are not grammatical errors but would not 'sound right' to most fluent

speakers of US English."). These labels provide a distinction between true errors and other types of disfluency, although this distinction is more subjective for transcribed speech (as in our data) compared to standard written text. In our final analysis, we collapse this distinction into a single category of **Fluency issues**. Slightly more than half of the segments (208 or 52.0%) were labeled as disfluent, which aligns with the 52.5% marked as fluent in the original annotation. However, there is a discrepancy where 20.2% (42) of examples labeled as having fluency issues were previously labeled as fluent, and 17.33% (35) of examples marked as disfluent in the original annotation were not labeled as having fluency issues in this annotation. There are no consistent trends in other features that could explain this discrepancy, suggesting that fluency may simply be somewhat subjective.

The **MT-specific issue** labels were: **Duplication** ("The translation contains repeated words or phrases unlikely to have been repeated in the original text."), **Non-words** ("Transliterations, untransliterated source language words, words constructed from parts of other words, or other non-words."), and **Bad copy** ("One or more tokens that should have been passed through (e.g., URL, email address, numbers) appear to have been changed or broken."). Only 7.75% of annotated segments (31) displayed one or more of these MT error labels. Only one segment was labeled with a Bad copy error. More than half of segments with duplication (52.63% or 10 segments) were originally labeled as believable compared to only 16.67% (2 segments) with non-words. This suggests that duplication is an error that users may be able to see past when it only affects fluency and not meaning, while non-words are an indicator of meaning loss.

Annotators were also given options to note when it appeared that content had been dropped based on context, when the result was humorous or explicit, and whether disfluencies sounded like mistakes a human, such as a language learner, might make. These labels proved too subjective to be useful in the analysis.

To distinguish between fluency and believability, we focus on the segments the original annotation found to be either fluent but not believable or believable but not fluent, as shown in Table 4.9. Unbelievable fluent translation have more issues overall, but especially plausibility issues (50% vs 9%). Believable disfluent translations have fewer intelligibility issues and plausibility issues.

	Any issues	Intelligibility	Plausibility	Fluency	MT-specific
Fluent					
+Believable	17.00%	2.00%	9.00%	10.00%	1.00%
+Unbelievable	70.41%	8.16%	50.00%	31.63%	7.14%
Disfluent					
+Believable	90.10%	6.93%	26.73%	81.19%	11.88%
+Unbelievable	94.06%	33.66%	70.30%	84.16%	10.89%

Table 4.9: Percent of segments with each category of issue that were labeled as each combination of believable and fluent.

4.6 Summary and Implications

This work used traditional NLP annotation methods to measure users' perceptions of *believability* of MT output and qualitative methods to identify additional factors that affect believability. These methods allow us to identify broad relationships between believability and traditional MT quality metrics, fluency and adequacy, showing that believability is strongly correlated with fluency. Qualitative

analysis shows that believability has elements of traditional fluency as well as plausibility. When output is so disfluent as to be incomprehensible, users cannot believe that the meaning of the output is correct because they cannot glean meaning from the output. When output is not semantically plausible, users are unlikely to believe that it is correct regardless of fluency.

These results suggest that users are most likely to be misled by inadequate MT outputs when they are comprehensible and plausible in context. An MT system that optimizes for fluency and cohesion could effectively mask errors, producing misleading output, but MT quality estimation (see, e.g., Specia et al. (2009)) could be used to flag potentially misleading output.

To truly understand the effects of potentially misleading MT in context, we need to look at a particular use case and see whether users are actually misled and whether we can help users with simple interventions.

Chapter 5: Mitigating Misleading MT in the Real World: Intelligence Analysis

5.1 Introduction

Having conducted experiments to explore the characteristics of potentially misleading machine translations in Chapters 3 and 4, the next step is to determine the impact of misleading MT in a specific use case and to explore practical interventions that may help mitigate the risks of misleading MT. The use case chosen for this research is *MT-enabled scanning* of foreign language text by Intelligence Analysts (IAs). Scanning is the process of triaging a collection of documents to identify those that are worth further analysis. Ideally, foreign language material would be scanned by *language-enabled* analysts: analysts fluent in the document language. In practice, however, there is too much foreign language material and too few language-enabled analysts, or language analysts, to scan the data in a timely fashion (Robb et al., 2005). Machine translation can enable IAs with little or no familiarity with the document language to assist in scanning, though it also raises the temptation to perform additional analytic tasks based on MT output without human translation.

The remainder of this chapter provides further background on the MT-enabled scanning use case in the broader context of intelligence analysis and addresses the potential risks posed by MT use in scanning and other analytic tasks. We then propose a two-part intervention designed to help analysts appropriately calibrate their trust in the MT output and conduct a user study to establish a baseline of the effects of potentially misleading MT output on analyst scanning performance and confidence and to demonstrate the impact of the interventions.

5.2 Intelligence Analysis

To understand the risks and rewards of using machine translation in the intelligence analysis process, we must first understand the typical process and the official standards analysts must follow. Intelligence Analysis is the process of "integrating, evaluating and analyzing data, and distilling it into final intelligence products." (Intelligence Community, 2015).

As a hypothetical example, imagine a team of analysts has received the contents of the chat server for the Zendian military's death ray development team, project 'HIGH NOON.' They first need to identify conversations that are relevant to the information needs of policymakers and other customers. This is the scanning phase. If the primary need is information about weapon development progress, relevant conversations may reference the project name, deadlines, or technical questions or might include attachments such as reports, diagrams, or video of a test. Although present in the data set, conversations discussing where to have lunch, detailing a

recent romantic encounter, or sharing memes would be of little or no relevance to the information need. The chats and attachments are likely to be in the Zendian language, and there may be only one or two language analysts on the team who are fluent in Zendian. When relevant conversations, attachments, and individual messages are identified, they are translated into English so that the non-language-enabled analysts can analyze and assimilate the key information to produce reports for dissemination to intelligence customers. Translations and reports are subject to one or more phases of quality review to maintain rigor and accuracy in the analytic process.

In response to recommendations from the 9/11 Commission, the Office of the Director of National Intelligence issued Intelligence Community Directive (ICD) 203, *Analytic Standards* (McConnell, 2007), to encourage more rigorous analysis. ICD 203 introduces four overarching standards for intelligence products (objectivity, independence from political consideration, timeliness, and based on all available sources) and a set of nine Analytic Tradecraft Standards, commonly referred to as the "Analytic Integrity Standards" or AIS. These standards require that each intelligence product:

1. Properly describes quality and credibility of underlying sources, data, and methodologies

2. Properly expresses and explains uncertainties associated with major analytic judgments

3. Properly distinguishes between underlying intelligence information and ana-

lysts' assumptions and judgments

4. Incorporates analysis of alternatives

5. Demonstrates customer relevance and addresses implications

6. Uses clear and logical argumentation

7. Explains change to or consistency of analytic judgments

8. Makes accurate judgments and assessments

9. Incorporates effective visual information where appropriate

These standards affect each stage of the process. In the Zendian example above, providing *timely* analysis *based on all available sources* requires scanning the document collection quickly, accurately, and thoroughly so as not to miss relevant documents. AIS #5 requires that scanning analysts understand the information needs to identify relevant documents. Analysis of the relevant documents relies on translation, which is performed using a methodology that gains credibility (AIS #1) by adhering to the AIS in the translation and quality review process itself. When translating relevant documents, language analysts may apply AIS #2 and #3 by adding analytic comments to denote places where the literal translation may differ from the intended meaning. In the translation quality review, the reviewer and translator may apply AIS #4 by discussing ambiguities and potential alternate translations. Analyzing and assimilating content across documents to "make accurate judgments and assessments" (AIS #8) requires a clear understanding of the

documents, context, and participants in the communication to assess the credibility (AIS #1) and enough understanding of prior intelligence reporting to know whether it indicates a change (AIS #7). AIS #6 and #9 are applied in the writing process, and the report is reviewed to ensure that the report is consistent with the source material and translation and that the AIS have been applied.

5.3 MT in Intelligence Analysis

Machine Translation could potentially be used at any stage of the analytic process where non-language-enabled analysts interact with foreign language material, but the risks and AIS implications differ depending on the use case and process. The primary ways IAs may use MT include scanning foreign language text and reporting directly off MT.

5.3.1 Scanning Use Case

In our scanning scenario, an Intelligence Analyst is tasked with responding to information needs within their area of analytic expertise based on source material, including foreign language text. The foreign language text we will call "documents" although the type and length of documents may vary from technical reports to short, informal messages. Analysts who do not know the language must rely on machine translation to get the gist of the documents and determine which are likely to contain relevant information and which are "NTR" (Nothing to Report). This is a *relevance judgment* task. The relevant documents are passed to language analysts

for translation, often with a contextual note explaining the information that the IA hopes to learn from the document. Generating the contextual note is a *comprehension* task, and the note may be general or specific depending on the analyst's level of understanding of the text and confidence in that understanding. Contextual notes are not required, but they provide insight into the scanning process and an opportunity for the language analyst to quickly skim the document to verify that it is relevant and reach back to verify in case of misunderstanding.

Potential scanning errors are: discarding a document with relevant information (false negative) and wasting language analyst time on irrelevant documents (false positive). The impact of the former is relative to the importance of the missed information: A single piece of crucial information missed could lead to broader intelligence failures. False negatives also have implications for one of the overarching standards, the requirement that intelligence products be based on all available sources. The impact of false positives is relative to the frequency of error: since every document is viewed by both an MT-enabled IA and a Language Analyst, passing too many irrelevant documents through may reduce the benefits of MT-enabled scanning, leaving language analysts overwhelmed.

Machine Translation quality and believability can lead to misleading translations that affect scanning decisions in several ways. When important words or phrases are untranslated, mistranslated, or hallucinated, it will be difficult for the analyst to make an accurate relevance judgment. Document context and analyst knowledge of the real-world context provide additional information but may not be enough to counter misleading translation. If the document contains other relevant

information, any missing or mistranslated information is not needed to justify labeling the document as relevant, and hallucinations that appear relevant will not affect the analyst's judgment. If the analyst has sufficient real-world context to infer missing information or identify mistranslated or hallucinated information, they will be able to make an accurate judgment rather than being misled by the MT output. Document context can play a similar role if the analyst finds the translation of the context believable enough to be willing to rely on that context. In both cases, the MT for the document as a whole would be inadequate but not misleading. However, when the analyst makes an incorrect relevance judgment due to a fluent but implausible mistranslation in the context of a believable mistranslation that overwhelms evidence from context, the MT output has misled the analyst.

MT output that is disfluent to the point of incomprehensibility lends itself to skimming for keywords rather than attempting to understand the text. When that skimming leads to misunderstanding, the analyst can be said to have been misled. In the Zendian example, an analyst who sees the word "noon" in the output may assume relevance to the "HIGH NOON" project in a conversation scheduling lunch or may see the word "lunch" and assume a conversation is irrelevant when a scientist is apologizing for missing lunch because they made a crucial breakthrough. The latter will likely never be reviewed, and the information will simply be lost. The former would be quickly identified as irrelevant, so it may be tempting to err on the side of caution, but when there are many such documents in the queue, it can add up to create a bottleneck in the analytic process.

Scanning errors may also occur because the text is difficult to judge. There

may not be enough context, or the source text itself may be written in a way that obfuscates or misleads. Because these issues are not specific to machine translation, they are not directly addressed in this work except to filter experimental data to remove these examples.

In sum, the circumstances in which MT output may mislead an analyst to make an incorrect relevance judgment are: (a) the MT output has omitted, mistranslated, or hallucinated key information in a way that is believable in context; (b) accurate translation of key information or context is not believable; or (c) the MT output lacks comprehensibility, obscuring context and leading the analyst to rely on keywords alone.

Although MT-enabled scanning risks missing relevant documents and may still result in more documents than the language analysts can process in a timely manner, these risks are primarily tied to the volume of content that can be analyzed: Without MT-enabled scanning by intelligence analysts, language analysts would be more overwhelmed and the likelihood of missing relevant documents due to lack of time would be increased.

5.3.2 Reporting Use Case

The second use case, reporting directly off MT output, poses more serious risks than scanning. Allowing analysts to report off MT rather than waiting for quality-reviewed human translation would increase the volume of content that is analyzed and reported, but it would add risks implicating AIS #1-4 and 8. Basing analytic

judgments on MT output adds ambiguity to questions of uncertainty and credibility (AIS #1-2), as the non-language-enabled analyst has no means to judge the accuracy of the MT output. Assuming that MT output is sufficient to understand the source text may indicate a failure to "properly distinguish between underlying intelligence information" (AIS #3) and a potentially inaccurate transformation like MT. Relying on a single MT output eliminates the ability to consider alternate translations in the human translation quality review process (AIS #4), and without quality review of a translation, MT errors may lead to inaccurate judgments and assessments (AIS #8).

To contrast these risks with the risks of MT in scanning, let us return to the Zendian hypothetical. If Zendian has a unique writing system and the MT output replaced some of the numbers and units of measurement with incorrect values in a death ray test report, the document would be correctly scanned as relevant despite the errors. If that document were passed along for human translation, the values would then be translated correctly, leading to correct reporting on the current capabilities of the death ray. But if the analyst saw the incorrect information and felt that it was such a large improvement in capability that it was too urgent to wait for human translation and immediately drafted an urgent report, the misinformation could have strategic, military, and/or diplomatic consequences.

Further, the MT need not be incorrect to be insufficient for reporting. Recall that the translation process may include analytic comments to clarify when the literal translation may differ from the intended meaning. If one of the Zendian engineers sent a message on the Zendian equivalent of April Fools' Day with an

attachment alleged to be a new requirements specification document, a perfectly accurate machine translation would not provide sufficient context to understand that the document is a parody. A language analyst might not pick up all of the jokes, but they would be more likely than a non-language-enabled analyst to recognize references to pop culture or classical literature and make the connection with the holiday, particularly with quality review by a second language analyst.

Despite the risks, the volume of foreign language data and the apparent quality of translation make it tempting to rely on MT without human translation, and at least one U.S. Government agency outside the intelligence community has even encouraged this use. In September 2019, ProPublica reported that U.S. Citizenship and Immigration Services officials had been encouraged to use Google Translate to translate social media posts in the process of vetting some refugees without subsequent human translation (Torbati, 2019). Although USCIS states that applicants will be given the opportunity to discuss social media posts that are deemed concerning (Hawkins, 2017), mistranslations may prejudice the official, and they may be disinclined to believe that the machine made a mistake even if the applicant can articulate the nature of the mistranslation.

Because the analysts in this study do not currently report directly off MT output, the reporting use case is not the primary focus of this study. However, addressing the comprehension element of the scanning use case (contextual notes) may provide some insight into the level of risk if analysts were to begin reporting directly off MT and whether or not the proposed interventions would reduce risk in this use case as well.

5.4 Proposed Interventions

This work proposes to test interventions to help analysts more accurately predict the relevance of documents translated by off-the-shelf MT engines. In this section, we establish the parameters we will use in choosing interventions, explore the space of potential interventions, and select two interventions that meet our requirements.

Recall that the scanning use case occurs in an environment where there is so much foreign language content across a range of topics and document types that it is not practical for language analysts alone to triage. Even the languages needed may change in response to world events or new intelligence priorities. Finite computational resources are available to support these needs, so resource-intensive interventions may not be practical. Expertise is often limited as well. The teams that integrate and deploy machine translation solutions for analysts are more likely to be generalist software engineers than machine learning engineers or machine translation experts, as these skills are in high demand. This, too, may limit the practicality of interventions. The ideal intervention does not require specialized skills to implement, is not domain-specific, and provides enough of a boost in analyst performance to justify any increase in computational resources needed.

Given the scenarios in which potentially misleading MT output may induce scanning errors detailed in Section 5.3.1, the goal of the interventions is to help analysts calibrate their judgments of MT output to decrease the believability of errors while increasing the believability of accurate translations. Without intervention, the

user has only document context and real-world context to apply in deciding whether the MT output is an accurate representation of the meaning of the source text. We will discuss three categories of approaches: Improving model quality, Scoring model output, and Providing alternate translations.

5.4.1 Improving Model Quality

One potential solution to reduce how often analysts are misled by MT output is to reduce how often the output itself is potentially misleading. Even with the early NMT models in Chapters 3 and 4, better models produced fewer fluently inadequate translations and fewer believably inadequate translations. However, improving models requires in-domain bitext which may not be available and expertise to ensure best practices are followed to prevent issues like catastrophic forgetting (e.g., Thompson et al., 2019; Gu and Feng, 2020; Cao et al., 2021; Carrión and Casacuberta, 2022; Shao and Feng, 2022) so the model can continue to be useful across the wide range of document domains and genres. Training or tuning models may also require reallocating computational resources away from the platforms MT is deployed for use by analysts. Even with the computational resources, data, and human expertise necessary to build and deploy improved MT models, improvement is not perfection, and simply deploying a better model does not provide a means for assisting analysts when the model makes mistakes. The ideal intervention can immediately be deployed with MT models of any quality and will have the potential to continue to help analysts even as newer, better models are deployed.

5.4.2 Scoring Model Output

Another potential solution is to provide confidence or quality estimates with translations. Well-calibrated scores would help analysts apply AIS #1 by allowing analysts to more accurately assess the quality of the MT element of their methodology. Unfortunately, as noted in Section 2.2.3, sentence-level confidence scores tend not to be well-calibrated and lack the specificity needed to help the user decide which parts of the translation to believe. Fine-grained error detection is a promising line of research, but the best models do not perform well enough and require considerable resources, with the top submissions 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).

5.4.3 Providing Alternate Translations

The best fine-grained error detection model at WMT'23 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? Anecdotally, analysts prefer MT interfaces that provide output from multiple MT systems. Over several years of delivering analyst training on effective MT use before NMT became widely available, multiple analysts told the author they believed the MT they saw in a user interface that displayed outputs from a commercial MT system and government-produced rule-based MT system side-by-side by default was more accurate than the MT in another user interface that showed only the same commercial MT output by

default but allowed the user to request the RBMT. Developers of the second interface chose the commercial MT as the default for languages for which they believed it would provide better translations than the RBMT system. Surprisingly, even after developers of the second user interface changed its default to a new higher quality NMT model, users of the first interface, which was still using the commercial MT from 2016, expressed concern that the tool was being decommissioned because they felt the translations were more accurate. The analysts' view of the combination as more accurate than just the commercial MT or even NMT suggests that what they perceive as the overall quality of the MT was based on not just the individual MT outputs but the totality of information the analysts gleaned from the combination of outputs. Whether analysts are actually able to triangulate more information from two MT outputs or are simply overconfident is a question we can test in our user study.

This type of intervention is appealing because it requires no additional data or specialized skills, can be used for any language where more than one MT system is available, and is already known to be preferred by some analysts. It also fits with AIS #4, "Incorporates analysis of alternatives," although it should be noted that the alternatives here are not necessarily alternate translations based on different interpretations but may include one or more outputs that are entirely disconnected from the input.

Inspired by the anecdote above, we propose two versions of this intervention and a combination of the two versions. First, we propose pairing output from a single NMT system with output from a second NMT system since these are the highest-

quality MT systems available. Second, we propose adding a rule-based MT (RBMT) output to one or two NMT outputs. These interventions and their expected impact on believability calibration to mitigate potentially misleading machine translation outputs are described in Sections 5.4.3.1 and 5.4.3.2

5.4.3.1 Intervention 1: Provide outputs from two NMT models

As discussed in Chapters 3 and 4, non-language-enabled Intelligence Analysts with output from one MT system can only rely on features of the output text like fluency and 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 would provide additional information to inform the decision. As noted in Section 2.3.2, 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). Similar improvements can be expected in the MT-enabled scanning use case. Because hallucinations are often tied to the training data (Raunak et al., 2021), we can expect that if the first MT output is believable but contains a hallucination that affects relevance judgment, output from a second model trained on different data is unlikely to produce the same hallucination. The difference between the two translations draws attention to the potential error, and the two contexts can inform the analyst's interpretation of the difference. A degenerate repetition or other disfluency would reduce the believability of a single MT output that is otherwise adequate, but if the second MT output has a similar

meaning, the adequate content will be raised in prominence because it appears in both and the analyst may choose to believe it more confidently as a result.

To be effective in this setting, the two MT models should be of competitive quality but sufficiently different in structure and/or training data to be unlikely to produce the same errors. If one model is significantly lower quality, it may be too unreliable to provide benefit, and if the models produce the same errors, those errors will be reinforced. Deploying a second NMT model approximately doubles the computational resources required for each translation, but if at least one of the models is multilingual, those resources can be shared across languages. To this end, we will use a freely available massively multilingual pre-trained model, NLLB-200 (Koishekenov et al., 2022), and a commercial off-the-shelf system, SYSTRAN. More information about these models and their applicability in this setting is in Section 5.5.7.

5.4.3.2 Intervention 2: Provide Rule-Based MT output

Rule-Based MT (RBMT) is not expected to provide the readability of neural MT output, but it does provide more interpretability than off-the-shelf NMT because every word or phrase in the output is a translation of specific words and phrases in the source. It is also easy to update with new named entities and specialized terminology. For these reasons, the CyberTrans MT platform available to analysts throughout the U.S. intelligence community includes Motrans RBMT for many languages (Martindale, 2012). Paired with one or more NMT outputs, Mo-

trans can provide a similar emphasizing effect to displaying a second NMT output. Paired with two NMT outputs with significant meaning differences, the Motrans output can serve as a "tie-breaker" by connecting back to the source. A third NMT could provide a similar benefit, but NMT systems require significantly more computational resources than Motrans, and the interpretability of Motrans could prove useful in itself. It is possible that the lack of fluency for Motrans could result in the NMT being consistently preferred, but if Motrans's interpretability increases believability enough to overcome the disfluency, it may still provide a useful check on the NMT output as it may have for the commercial MT in the two-MT interface in the anecdote.

It is worth noting that there are other ways to achieve a similar connection between output and source text. An NMT system can be modified to provide alignments (e.g., Ding et al., 2019), but this requires a deeper understanding of NMT architecture than can be expected of a typical integrator of off-the-shelf MT. Statistical MT output can be aligned with input in a more straightforward way than NMT, but SMT models require much more memory than Motrans, and off-the-shelf SMT models are no longer readily available. These options also lack Motrans's ability to rapidly add named entities and specialized terminology.

5.5 User Study Design

Recall from the previous section that the goal of our interventions is to mitigate potentially misleading machine translations by helping users to find otherwise

believably inadequate translations less believable and otherwise unbelievably adequate translations more believable, preventing them from being misled. Rather than directly test how these interventions affect the MT output's believability in a vacuum, we seek to provide a realistic context that mirrors analyst scanning tasks. We infer the effectiveness of the interventions by measuring the performance of actual analysts on these tasks. For the purposes of this study, we consider an analyst to have been misled by the MT output if they make an incorrect relevance judgment when scanning or if they provide inaccurate information in a contextual note where a language analyst reading the source text would not make a similar mistake. Analyst accuracy ties back to AIS #8, "Makes accurate judgments and assessments." We will also measure the interventions' effects on analyst confidence to help answer whether the perception of higher accuracy in the anecdote in Section 5.4.3 was a reflection of overconfidence or was justified by analysts' ability to glean more information from the translations. Analyst confidence also implicates AIS #3 as confidence calibration reflects analysts's ability to distinguish between their judgments based on the MT output and the actual content of the source text.

Key research questions for the user study are:

1. How accurately can analysts scan short documents at different levels of comprehension based on output from one NMT system? How confident are analysts in those judgments, and is that confidence justified by their accuracy?

2. Does providing output from a second NMT system increase their accuracy? Does it help the analyst appropriately calibrate their confidence?

3. Does adding rule-based MT output affect analyst accuracy and/or confidence?

The levels of comprehension tie back to the use cases in Section 5.3. Relevance judgment is the minimum comprehension necessary for the scanning use case. The analyst need not fully understand the content of the document, only that the document is relevant enough to want to understand the content after human translation. This decision is typically reflected by assigning a disposition to the document indicating the next step to be taken for relevant documents (in this case, translation) or "Nothing to Report" (NTR) for irrelevant documents. The construction of a contextual note for relevant documents may require deeper levels of comprehension if the analyst chooses to include more information. In this user study, relevance judgment accuracy serves as our primary means of measuring the impact of the interventions, but we use guided contextual note construction to provide further insight into the analysts' comprehension. Accuracy of contextual note information at deeper comprehension levels can provide insight into whether the proposed interventions improve comprehension enough to make MT a viable part of other phases of analysis.

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. This will address the first and second research questions. The within-subjects variable is whether the analyst is provided rule-based MT output from Motrans in addition to the NMT output(s). This will address the third research question.

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 with and without Motrans conditions 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 platform used to conduct the user study tracked the median time on task for each participant, allowing us to compare time between subjects in the one-NMT and two-NMT conditions but not within subjects to compare with and without RBMT. We report these results in Section 5.6.5, but caution that the user study environment may invalidate these metrics as analysts participated from their own workstations in their typical environment and could be interrupted by other urgent demands at any time. This reflected the typical environment for scanning tasks and made participating easier for analysts to justify, but it limits the conclusions we can draw about the effects of the interventions on scanning time.

5.5.1 Participants

The participants in the user study were civilian employees of a U.S. intelligence agency working in the Washington, D.C. area with significant experience (at least three months) performing scanning tasks with the aid of machine translation and little or no knowledge of the source language. Participants were recruited through messages on internal networking sites and mailing lists, with the goal of recruiting up

to 40 qualified analysts. Analysts were screened using a qualification survey, which also gathered relevant background information for qualified participants, including foreign language experience, intelligence analysis experience, familiarity with MT in general, familiarity with the technology underlying MT, and perceptions of the reliability of MT. When analysts 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 to ensure a counterbalance of ordering effects as described in Section 5.5. In total, 35 analysts responded to the survey, and 26 completed the user study. Two of the survey respondents did not qualify because their machine translation use was in a language they knew, and seven analysts did not respond to contact after the background survey. Two of the remaining 28 analysts, 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.

Regarding past experience with MT, all participants said they had used a standalone MT service on the internal network, and all but one had used MT integrated into other analyst tools. 85% said they had manually set basic translation options such as language or translation engine in the standalone tool, and 77% had changed MT settings in other tools. The majority of participants (58%) had only six months to one year of on-the-job experience using MT for scanning, but 35% had two to five years of experience, and two participants had more than five years of experience. Six participants (23%) listed Persian as one of the primary MT languages they had

used.

All participants self-declared English to be one of their primary communication languages for most of their adult lives, and all but one had used English since childhood. None of the participants spoke Persian Farsi, but four participants spoke another language that uses Perso-Arabic script. The source text in the tasks for these four users was obfuscated by a simple one-to-one substitution with unrelated Cyrillic characters. This substitution allowed users to still recognize if a comment referred to a previous commenter by their handle or if the MT output was significantly shorter or longer than the source text, but it prevented them from responding based on recognizing keywords in the source text instead of relying on the MT output.

5.5.2 Scenario

This study will focus 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 and linguistic differences with English. Despite these challenges, 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.

Although intelligence analysts interact with a variety of lengths and genres of documents, this study focuses on conversation threads. Messages in conversation threads are particularly difficult for relevance judgment because understanding any

given message requires understanding its context in the conversation. Conversational text also often uses colloquial language which may be out of domain for MT systems. For these reasons, performance on conversation threads may be seen as a lower bound on analyst relevance judgment performance more broadly. 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.

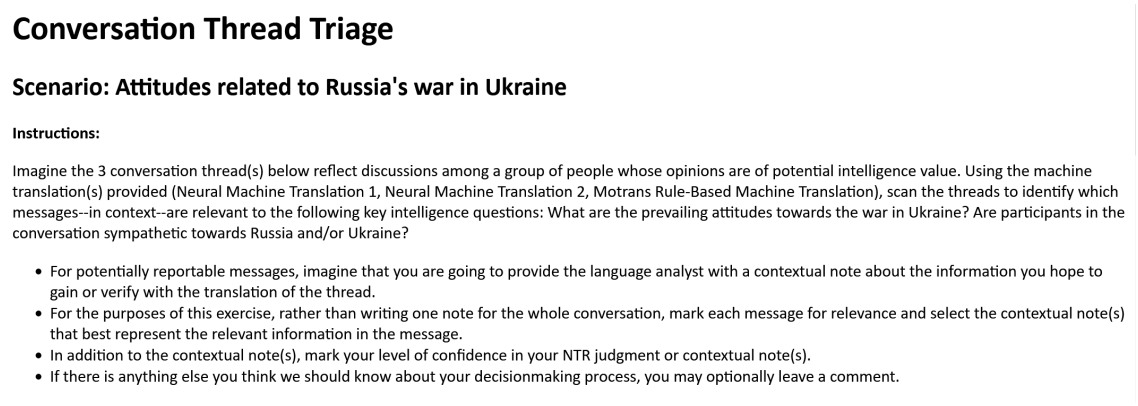
These topics were chosen because they are related to U.S. intelligence priorities (strategic competition and violent extremist organizations) and are likely to elicit reactions among readers of Iranian news articles because of Iran's support of Russia (Bowen et al., 2022) and Hezbollah (Humud, 2023) and Iran's stance against ISIS (Arango and Erdbrink, 2014).

We frame the scenario for the user study as follows:

> *Imagine the conversation threads below reflect discussion among a group of people whose opinions are of potential intelligence value. Scan the threads to identify which messages–in context–are relevant to the following key intelligence questions.*

For the Russia-Ukraine topic, the key intelligence questions (KIQs) are: "What are the prevailing attitudes with regards to the war in Ukraine? Are participants in the conversation sympathetic towards Russia and/or Ukraine?" For the terrorism topic, the KIQs are: "What are the prevailing attitudes towards known terrorist organizations? Are participants in the conversation sympathetic towards Hezbollah and/or ISIS?"

Figures 5.1 and 5.2 are screenshots showing how the scenario was displayed to users for each topic. The next section details how this scenario was implemented into tasks for the user study.



Conversation Thread Triage

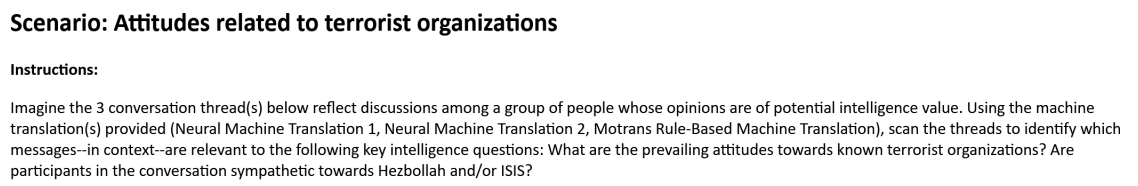
Scenario: Attitudes related to Russia's war in Ukraine

Instructions:

Imagine the 3 conversation thread(s) below reflect discussions among a group of people whose opinions are of potential intelligence value. Using the machine translation(s) provided (Neural Machine Translation 1, Neural Machine Translation 2, Motrans Rule-Based Machine Translation), scan the threads to identify which messages--in context--are relevant to the following key intelligence questions: What are the prevailing attitudes towards the war in Ukraine? Are participants in the conversation sympathetic towards Russia and/or Ukraine?

- For potentially reportable messages, imagine that you are going to provide the language analyst with a contextual note about the information you hope to gain or verify with the translation of the thread.
- For the purposes of this exercise, rather than writing one note for the whole conversation, mark each message for relevance and select the contextual note(s) that best represent the relevant information in the message.
- In addition to the contextual note(s), mark your level of confidence in your NTR judgment or contextual note(s).
- If there is anything else you think we should know about your decisionmaking process, you may optionally leave a comment.

Figure 5.1: Screenshot of the heading of a Russia/Ukraine task.



Scenario: Attitudes related to terrorist organizations

Instructions:

Imagine the 3 conversation thread(s) below reflect discussions among a group of people whose opinions are of potential intelligence value. Using the machine translation(s) provided (Neural Machine Translation 1, Neural Machine Translation 2, Motrans Rule-Based Machine Translation), scan the threads to identify which messages--in context--are relevant to the following key intelligence questions: What are the prevailing attitudes towards known terrorist organizations? Are participants in the conversation sympathetic towards Hezbollah and/or ISIS?

Figure 5.2: Screenshot of the scenario section of a Hezbollah/ISIS task.

5.5.3 Tasks

Analogous to the real MT-enabled scanning 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 (KIQs). 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 provide confidence scores for their judgments. A screenshot of the interface for a task is shown in Figure 5.3.

The contextual note information was gathered in a multiple-choice, fill-in-the-blank style, as shown in Figure 5.4. 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 a given entity. Analysts could reflect 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 reflect uncertainty about the stance of 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.

The study was conducted on a platform that tracks time on task, and participants were encouraged to complete each task in one sitting without interruption.

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 *PLEASE CHOOSE "NTR" OR "RELEVANT" ABOVE*

Figure 5.3: The user interface for the relevance judgment step of a Hezbollah/ISIS conversation thread.

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 How confident are you that this post is NTR? ○ no confidence ○ somewhat confident ◉ mostly confident ○ confident ○ almost certain Comments (optional):
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 5.4: The user interface for the contextual note step of a Hezbollah/ISIS conversation thread.

5.5.4 Qualitative Measures

As noted in Section 5.5.1, analysts were screened using a qualification survey, which also gathered information about their perceptions of the reliability of the MT they typically use for scanning foreign language text in a language they do not understand. After finishing all four user study tasks, they completed a post-task survey, which included questions related to the difficulty of the task and how similar it was to their typical scanning tasks. The post-task survey also asked about their user interface preferences and level of trust toward the NMT and Motrans outputs in the user study and provided space for open-ended feedback about their experiences with machine translation on the job and in the user study. This feedback provides a window into participant perceptions of the individual MT systems in the user study as well as combinations.

To observe user trust, we lightly modify the trust scale recommended in Hoffman et al. (2018), which combines and adapts questions that have previously been shown to be highly reliable. In the pre-task survey, participants responded to the statements with the context of, "Thinking about the machine translations you see in the tools you use when scanning text in a language you do not know, how much do you agree with the following statements?" In the post-task survey, they were asked about each of the MT systems they saw individually and in combination, except for the consistency question, which was only asked about individual MT systems because combining them does not affect their consistency. The other statements were modified slightly for the combinations (e.g., "I could count on at least one of them

to be correct" and "With both NMT1 and NMT2 together, I was confident...").

5.5.5 Corpus Collection

The initial corpus of comment threads was collected by searching Persian-language news sites[1] for keywords related to topics of interest and then scraping the user comments and their visible metadata (username, timestamp, and threading information) from the articles that were returned. Keywords used were: روسیه (Russia), اوکراین (Ukraine), داعشی (ISIS), and حزب‌الله (Hezbollah). This resulted in 5,917 comments for the Russia-Ukraine topic and 2,238 comments for the terrorist groups topic. For purposes of this work, a conversation thread is defined as a comment and its replies. To increase the chance of a thread having some back-and-forth, we use only threads with at least two replies, reducing the collection to 1,552 comments in 315 threads for Russia-Ukraine and 346 comments in 82 threads for terrorism.

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. To automatically identify these comments, we focused on the stance element of the task and ran the Twitter-trained sentiment analysis model from TimeLMs (Loureiro et al., 2022) on the machine translations of the comments and chose 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.

[1] isna.ir/news, tabnak.ir/fa/news, khabaronline.ir/news

Figure 5.5: Screenshot of the relevance judgment and MT quality annotation view.

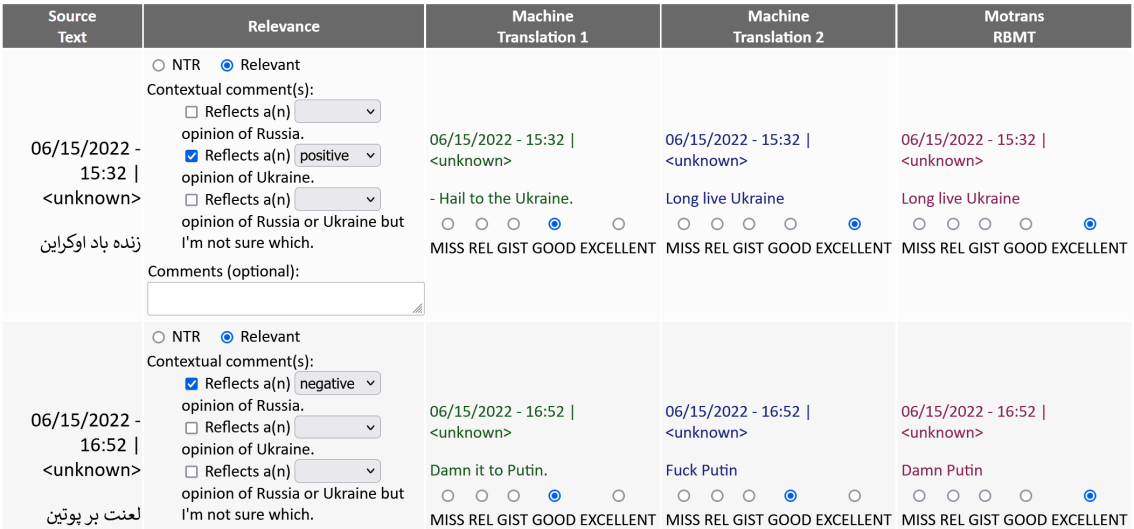


Figure 5.6: Screenshot of a completed comment annotation.

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

5.5.6 Annotation

Two Persian Language Analysts were recruited to provide gold standard annotations on the 210 Russia-Ukraine comments and 346 terrorism comments. The annotators completed the same relevance judgment 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 outputs, as discussed in more detail in Section 5.5.7.2. The interface also provided a comment box for each relevance judgment to allow annotators to note if there were issues that might affect whether the comment would be a useful example for the user study.

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 5.1. 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.

5.5.7 Machine Translation Models

To preserve generalizability across intelligence agencies, we have chosen models based on their applicability to the user study scenario rather than choosing based solely on MT models available at the agency where the user study was conducted. The NMT models chosen for this task are NLLB-200 and SYSTRAN. 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, including languages of strategic importance like Persian. SYSTRAN is familiar to U.S. government users through a long-standing collaboration with the Air Force (SYSTRAN, 2021) and previous integration in government translation platforms such as CyberTrans (Reeder, 2000). Because NLLB-200 is a massively multilingual model while SYSTRAN uses separate models for each language pair, we expect that the difference in training context and data will likely lead the systems to make different errors since hallucinations are often tied to the training data (Raunak et al., 2021). To test the NMT models' performance, we conducted automated evaluations on a standard test set as well as

the human evaluation included in the annotation phase.

The rule-based MT system, Motrans, 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 untranslated tokens into smaller translatable words with '+' between the resulting translations in the output. The Motrans translations of the second comment in Figure 5.3 in Section 5.5.3 demonstrate both these features. The ambiguous Farsi word کی is translated as *who/when*, and the incorrectly spaced phrase زنده بادحزب اله is translated as *long live+Hizbollah*. Automatic metrics do not distinguish between fluency and adequacy errors and cannot account for the alternate translations, so automatic scores for Motrans would be neither informative nor comparable with scores for the NMT models and are not provided in this work.

5.5.7.1 Automated Evaluation

Model	BLEU	chrF2++
ParsiNLU	23.3 ± 0.9	49.4 ± 0.7
M2M-100	27.0 ± 1.0	53.7 ± 0.7
NLLB-200	**33.3 ± 1.0**	**57.7 ± 0.8**
SYSTRAN	27.1 ± 1.0	53.2 ± 0.7

Table 5.2: Performance on the FLORES-200 devtest set for SYSTRAN version SPNS 9.7 and three pre-trained models: persiannlp/mt5-large-parsinlu-opus-translation_fa_en (ParsiNLU), facebook/m2m100_418M (m2m-100), facebook/nllb-200-distilled-600M (NLLB-200).

As shown in table 5.2, both models perform competitively on the FLORES-200 benchmark (NLLB Team et al., 2022). FLORES-200 is the only publicly available test set for Persian-English translation that is unlikely to overlap with training data for the MT systems being tested because it uses new translations rather than harvesting found bitext. The table compares the results of three pre-trained models available through the Huggingface model collection[2] and the version of SYSTRAN used in this work, SPNS 9.7. The ParsiNLU model (Khashabi et al., 2021) serves as a baseline pre-trained single language pair model, and M2M-100 serves as a baseline pre-trained multilingual model. NLLB-200 outperforms all three other models in both BLEU[3] and chrF2++[4] measured with SacreBLEU (Post, 2018). SYSTRAN and M2M-100 achieved almost the same scores and performed much better than ParsiNLU.

Because the colloquial style of the comment threads is different from the more standard register of the source material for FLORES-200 (Goyal et al., 2022), performance on FLORES-200 may not accurately reflect performance on the data for this task. However, in a scenario where an IT team is selecting an off-the-shelf MT system to integrate into analyst tools, they may not have access to in-domain test data or may not realize the need for in-domain test data and will likely accept results on available benchmarks. Based on this scenario, we select NLLB-200 as the primary MT model (NMT1) and SYSTRAN as a contrastive model (NMT2) to demonstrate Intervention 1.

[2]https://huggingface.co/models
[3]Signature: nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0
[4]Signature: nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.1.0

5.5.7.2 Human Evaluation

To test whether the differences in training contexts and data lead to different errors and to gain an understanding of the models' in-domain performance, we conducted a human evaluation as part of the annotation described in Section 5.5.6. 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 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.

Tables 5.3 shows the percent of translations from each MT system that were given each of the labels and the percent that were potentially misleading (below

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 5.3: Human 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 5.4: Annotator agreement on MT quality labels (Kendall's tau) for our two annotators on the 556 comments (210 Russia/Ukraine; 346 Hezbollah/ISIS).

GIST). Table 5.4 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.

5.5.8 Example Selection

From the annotated conversations, we selected 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 of the comments 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. Although the comment threads were chosen based on the presence of unambiguous relevance judgments, other comments in a selected thread may have had ambiguous relevance, entity, or sentiment labels. Any ambiguous labels in the selected threads were resolved through a tie-breaker annotation as noted in Section 5.5.6.

The distribution of relevance, entity, and sentiment labels for comments in each task is shown in Table 5.5. 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

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

Table 5.5: Distribution of gold standard relevance, entity, and sentiment labels for comments chosen for each task. *Pres* indicates how often that entity was judged to be a relevant entity, and *POS* and *NEG* indicate how often the sentiment towards that entity was positive or negative, respectively.

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.

To counterbalance for these distributional differences, we ensure that every participant sees every task, and we report aggregated relevance judgment performance across all tasks and aggregated entity and sentiment label performance across entities.

The combinations of MT quality labels are shown in Table 5.6. The MT

	NMT1		NMT2		Motrans	
	<GIST	GIST+	<GIST	GIST+	<GIST	GIST+
Russia/Ukraine A	56.3%	43.8%	43.8%	56.3%	56.3%	43.8%
Russia/Ukraine B	35.7%	64.3%	14.3%	85.7%	35.7%	64.3%
Hezbollah/ISIS A	26.7%	73.3%	46.7%	53.3%	20.0%	80.0%
Hezbollah/ISIS B	43.8%	56.3%	25.0%	75.0%	50.0%	50.0%
All	41.0%	59.0%	32.8%	67.2%	41.0%	59.0%

Table 5.6: MT quality labels for NMT1, NMT2, and Motrans in each task.

quality labels have been collapsed into <GIST (MISS or REL-ONLY) and GIST+ (GIST or better) to reflect the distinction between potentially misleading examples and non-misleading examples. Across all tasks, 41% of NMT1 translations are potentially misleading. NMT1 translations are key because NMT1 translations are shown to all users, and providing a second NMT output or RBMT output may help most when NMT1 translations are potentially misleading. Table 5.7 shows the percentage of comments with various combinations of MT quality. All MT outputs are potentially misleading for 11.5% of comments and above GIST quality for 31.1% of comments. When NMT1 is potentially misleading, NMT2 is also potentially misleading in 19.7% of comments, and Motrans is potentially misleading in 19.7% of comments. When NMT1 is GIST or better, NMT2 is potentially misleading in only 13.1% of comments, and Motrans is potentially misleading in 21.3% of comments. The higher percentages of potentially misleading translations when NMT1 is also potentially misleading suggest that these comments may be more difficult to translate and perhaps more difficult for users to judge.

		NMT2		Motrans	
		<GIST	GIST+	<GIST	GIST+
NMT1	<GIST	19.7%	21.3%	19.7%	21.3%
	GIST+	13.1%	45.9%	21.3%	37.7%
		All three <GIST		**All three GIST+**	
		11.5%		31.1%	

Table 5.7: Combinations of NMT1 quality paired with NMT2 and Motrans quality.

5.6 User Study Results

In discussing the results of the user study as they apply to the research questions from Section 5.5, we will first describe feedback from participants on how well the user study scenario reflected their typical triage tasks based on the pre- and post-task surveys. We will then analyze the quantitative results of the user study. Finally, we will share the qualitative results from analyst feedback in the post-task survey and briefly discuss issues related to time on task.

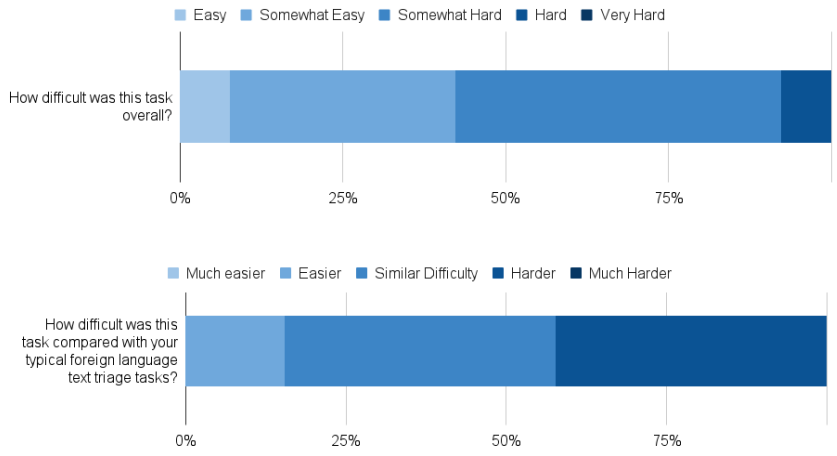


Figure 5.7: Responses to survey questions related to the difficulty of the user study.

5.6.1 Scenario Validation

The post-task survey included questions about how the user study tasks compared to participants' experience on the job. Figure 5.7 shows the responses to three questions related to the difficulty of the user study. The first question was, "How difficult was this task overall?" Slightly more than half of users (58%) found the task to be "Hard" or "Somewhat Hard." The second question compared the user study to the participants' "typical foreign language text triage tasks." 15% of users said it was easier, 42% said it was of similar difficulty, and 42% said it was harder. No participants said it was "Much Easier" or "Much Harder." These responses suggest that the tasks in the user study were reasonably challenging and generally similar difficulty to the more challenging intelligence analysis foreign language triage tasks.

Participants also provided open-ended responses to questions about what elements of the user study were similar to or different from their typical foreign language text triage tasks. Other than the use of machine translation in general, the most frequently mentioned similarity was the overall framing, mentioned by 11 analysts. Five analysts mentioned similar MT quality, and four analysts mentioned the difficulty of the task. One analyst's comment was difficult to categorize between the difficulty of the task and the quality of the MT output. The comment was, "Both had similar amounts of emotional outbursts." Three analysts noted similarities in the type of text, noting that the conversations lacked context and that the comments were short and informal. However, two analysts cited the length of text as a difference, noting that they typically triage whole documents. Other differences

that were mentioned included language (three analysts), topic (seven analysts), and familiarity with the topic (five analysts). Only one analyst mentioned a difference in the structure of the task, stating that they don't typically write contextual notes, "but its [sic] a good idea."

The availability of multiple MT outputs appeared in seven comments about similarities and six comments about differences. The seeming contradiction likely stems from differences in the availability of MT outputs for different languages and in different tools. The standalone MT service provides two MT outputs side-by-side by default for some languages, while other languages require the user to change display options to see more than one output, and other languages only have one MT system available. None of the other analyst tools that have integrated MT provide side-by-side output from more than one MT system.

Overall, the analyst comments about similarities and differences suggest that the scenario for the user study is comparable to many analyst workflows. Based on this feedback, we believe that the scenario framing this user study and the conversation threads selected for inclusion are analogous to at least some real-world scanning use cases.

5.6.2 Quantitative Analysis Methods

As described in Section 5.5.3, the user's response to each comment provides a confidence score and 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. For a given news comment, if the user's label of "Relevant" or "NTR" matched the gold standard, they received an accuracy score of 1 for relevance for that item and 0 for the item if it did not match. For entity accuracy, the user could indicate the presence of an opinion related to one or both of the entities or state that they believed there was an opinion about one of the entities, but they could not tell which. Rather than assign an arbitrary value to the unsure label, we can treat the accuracy as an ordinal scale with "incorrect" as the lowest score for missing both entities, "correct" as the highest score for getting both entities right, and "iffy" for responses that missed one entity or were unsure. Sentiment accuracy is similar but adds another level of complexity by allowing the user to state that they are unsure about the sentiment in addition to being unsure about which entity. We use the same ordinal categories for sentiment as for entity. We add responses where the user is unsure about both entity and sentiment to the "incorrect" bin, along with responses where the user is incorrect about one entity and unsure about the other. We add responses where the user is correct about one entity and unsure about the other to the "correct" bin.

We want to see whether NMT, 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 for user and item. The formula for the relevance model is:

$$y = \beta_0 + \beta_1 N + \beta_2 R + \beta_3 N * R + U + I + \epsilon \tag{5.1}$$

Where N is the presence of a second NMT, R is the presence of RBMT, U is a random intercept to control for individual user differences, and I is a random intercept to control for differences in difficulty between specific items. The β terms are the model parameters from which we can get the odds ratio of each effect on accuracy (right or wrong), and ϵ is an error term.

For the ordinal entity and sentiment accuracy levels, we use cumulative link mixed effects models (CLMM), an implementation of ordered or ordinal regression (McCullagh, 1980) from the ordinal R package[5]. The formula is similar to the standard GLMM formula but breaks y into ordered levels with thresholds:

$$y^* = \theta_j - (\beta_1 N + \beta_2 R + \beta_3 N * R + U + I) \tag{5.2}$$

In this case, y can be thought of as a set of $j + 1$ ordered bins corresponding to the levels of the response variable. Instead of a single β_0, we have a distinct θ_j for each threshold j between levels of the response variable. The β terms serve a similar purpose here as above, except that the odds are now for moving between adjacent levels. Note that ordinal regression models assume proportional odds, that is, that the fixed effects (NMT, RBMT, and their interaction) have the same effect at each of the levels of the response variable. In terms of entity and sentiment accuracy, that means we assume that the presence of RBMT, a second NMT, or both has the

[5]https://cran.r-project.org/web/packages/ordinal/index.html

same amount of effect on the odds of being more accurate in entity or sentiment judgments, whether starting from a baseline of being incorrect about both entities or only one.

To assess the impact on analyst confidence of adding RBMT and/or a second NMT, we fit four additional models. Following the pattern of the previous models, we fit a CLMM model with confidence as the response variable and NMT, RBMT, and their interaction as fixed variables. We also added relevance accuracy, entity accuracy, and sentiment accuracy as fixed variables. The proportional odds assumption in this model for confidence does not assume that it is equally easy to move between confidence levels 1 and 2 as it is to move between confidence levels 4 and 5, but it assumes that the impact (adding to the log odds or multiplying the odds) of the intervention or accuracy is the same. This model shows us the effects of the interventions on user confidence and whether those effects are mediated by the user's accuracy.

Alone, the confidence model does not provide enough information to tell whether user confidence is well calibrated with accuracy. To better understand whether confidence is a predictor of accuracy, we also fit models that replicate the relevance accuracy, entity accuracy, and sentiment accuracy models with confidence as a fixed variable. If confidence is a reliable predictor of accuracy, we will see a significant effect from confidence, and we will see improved model performance reflected in lower Akaike Information Criterion (AIC) values (Akaike, 1998) and higher likelihood ratios.

In addition to the fixed effects for user and item, we investigated including

another fixed effect for whether or not the quality of the first NMT output was labeled as potentially misleading. This quality effect was not statistically significant in any of the models, and the AIC was slightly higher for the models with quality. This indicates that including quality as a fixed effect does not improve the model's performance, so we do not include the quality effect in the models presented here (with or without confidence).

Noting that each regression inherently involves multiple hypothesis tests (one per fixed effect) and that we report results of multiple regression models, increasing the likelihood that a significant effect will be observed by chance (Type I error), we report both the raw p-values and p-values with False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) applied. Because we designed the experiment with the intent to assess the impact of the interventions on each of relevance accuracy, entity accuracy, sentiment accuracy, and confidence as separate analyses, we apply separate FDR corrections across all hypothesis tests for the same dependent variable. In other words, for each of relevance, entity, and sentiment accuracy, we apply an FDR correction to the nine p-values across the baseline model (four hypothesis tests) and the model with a fixed effect for confidence (five hypothesis tests). We also separately apply FDR to the p-values for the model with confidence as the dependent variable (six hypothesis tests).

5.6.3 Quantitative Results

In this section, we will discuss the results for each of Relevance, Entity, and Sentiment accuracy to address the analyst performance elements of the research questions. We will then turn to analyst confidence and confidence calibration.

5.6.3.1 Relevance Accuracy

Results of the GLMM with Relevance Accuracy as the response variable are shown in Table 5.8. 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 lower than just adding the second NMT.

Coefficient	β	Odds Ratio	Confidence Interval	$p_{uncorrected}$	p_{FDR}
(Intercept)	1.497	4.469	2.476 - 8.067	< 0.001	< 0.001
2-NMT	0.816	2.262	1.370 - 3.735	0.0014	0.0032
w/ RBMT	0.416	1.516	1.018 - 2.259	0.0408	0.0408
2-NMT+RBMT	-0.660	0.517	0.294 - 0.909	0.0219	0.0329

Table 5.8: Results of GLMM with Relevance Accuracy as the response variable (Number of observations: 1586, groups: item, 61; user, 26).

Coefficient	β	$exp(\beta)$	Confidence Interval	$p_{uncorrected}$	p_{FDR}
2-NMT	0.479	1.614	0.719 - 3.627	0.3627	0.4232
w/ RBMT	0.554	1.740	1.204 - 2.514	0.0032	0.0089
2-NMT+RBMT	-0.505	0.603	0.362 - 1.007	0.0530	0.0742

Table 5.9: Results of CLMM with Entity Accuracy as the response variable (Number of observations: 1586, groups: item, 61; user, 26).

5.6.3.2 Entity Accuracy

For entity accuracy (Table 5.9), 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.

5.6.3.3 Sentiment Accuracy

For sentiment accuracy (Table 5.10), 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.

5.6.3.4 Confidence

As shown in Table 5.11, 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), but after FDR correction (p=0.095) the increase is not statistically significant (p > 0.05). Nevertheless, the confidence interval suggests a tendency towards an improvement in user confidence with the addition of a second NMT output. Adding

Coefficient	β	$exp(\beta)$	Confidence Interval	$p_{uncorrected}$	p_{FDR}
2-NMT	0.456	1.578	0.433 - 1.485	0.1540	0.1848
w/ RBMT	0.251	1.285	0.933 - 1.770	0.1250	0.1848
2-NMT+RBMT	-0.331	0.718	0.460 - 1.122	0.1460	0.1848

Table 5.10: Results of CLMM with Sentiment Accuracy as the response variable (Number of observations: 1586, groups: item, 61; user, 26).

Coefficient	β	$exp(\beta)$	Confidence Interval	$p_{uncorrected}$	p_{FDR}
2-NMT	1.052	2.863	1.098 - 7.466	0.0315	0.0945
w/ RBMT	0.037	1.038	0.780 - 1.382	0.7980	0.7980
2-NMT+RBMT	0.098	1.103	0.755 - 1.612	0.6130	0.7356
Relevance Accuracy	0.349	1.417	0.924 - 2.174	0.1100	0.2200
Entity Accuracy	-0.404	0.667	0.264 - 1.691	0.3940	0.5910
Sentiment Accuracy	1.076	2.931	1.518 - 5.660	0.0014	0.0082

Table 5.11: Results of CLMM with Confidence as the response variable (Number of observations: 1586, groups: item, 61; user, 26).

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.

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 analyst confidence to the relevance, entity, and sentiment accuracy models and comparing the results.

With confidence as a fixed effect, we see minimal change in the effect of RBMT as shown in Table 5.12, 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 confidence is also a significantly better model based on AIC and log-likelihood (See Table 5.13). The large confidence

Coefficient	β	Odds Ratio	Confidence Interval	$p_{uncorrected}$	p_{FDR}
(Intercept)	1.665	5.287	2.930 - 9.545	3.2e-8	2.9e-7
2-NMT	0.577	1.781	1.053 - 3.013	0.0315	0.0362
w/ RBMT	0.445	1.560	1.038 - 2.345	0.0322	0.0362
2-NMT+RBMT	-0.685	0.504	0.284 - 0.895	0.0194	0.033
Confidence	1.176	3.241	1.898 - 5.534	1.7e-5	5.0e-5

Table 5.12: Results of GLMM with Relevance Accuracy as the response variable and confidence as a fixed effect (Number of observations: 1586, groups: item, 61; user, 26).

effect and model improvement suggest that analyst confidence is well-calibrated to relevance accuracy.

	AIC	Log Likelihood	χ^2	df	$p(>\chi^2)$
w/o Confidence	1335.8	-661.92	33.04	4	1.2e-6
w/ Confidence	1310.8	-645.39			

Table 5.13: Model comparison between Relevance Accuracy GLMMs with and without confidence as a fixed effect (Number of observations: 1586, groups: item, 61; user, 26).

Adding confidence to the entity model (Table 5.14) 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 demonstrates similar improvements in AIC and log-likelihood (Table 5.15) to those observed in the relevance model with confidence, 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 5.16). Confidence is the only

Coefficient	β	$exp(\beta)$	Confidence Interval	$p_{uncorrected}$	p_{FDR}
2-NMT	0.283	1.327	0.602 - 2.924	0.4823	0.4823
w/ RBMT	0.549	1.732	1.194 - 2.512	0.0038	0.0089
2-NMT+RBMT	1.181	3.258	1.194 - 5.409	4.9e-6	3.4e-5
Confidence	-0.514	0.597	0.357 - 1.001	0.0503	0.0742

Table 5.14: Results of CLMM with Entity Accuracy as the response variable and Confidence as a fixed effect (Number of observations: 1586, groups: item, 61; user, 26).

	AIC	Log Likelihood	χ^2	df	$p(>\chi^2)$
w/o Confidence	1807.6	-896.80	28.24	4	1.1e-5
w/ Confidence	1787.4	-882.68			

Table 5.15: Model comparison between Entity Accuracy CLMMs with and without confidence as a fixed effect

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 improvements to the AIC and log-likelihood of the model (Table 5.17) 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.

Coefficient	β	$exp(\beta)$	Confidence Interval	$p_{uncorrected}$	p_{FDR}
2-NMT	0.246	1.278	0.707 - 2.311	0.4162	0.4162
w/ RBMT	0.233	1.263	0.913 - 1.746	0.1584	0.1848
2-NMT+RBMT	-0.353	0.702	0.448 - 1.101	0.1233	0.1848
Confidence	1.321	3.748	2.420 - 5.800	3.2e-9	2.2e-8

Table 5.16: Results of CLMM with Sentiment Accuracy as the response variable and Confidence as a fixed effect (Number of observations: 1586, groups: item, 61; user, 26).

	AIC	Log Likelihood	χ^2	df	$p(>\chi^2)$
w/o Confidence	2507.2	-1246.6	48.89	4	2.8e-8
w/ Confidence	2474.3	-1226.2			

Table 5.17: Model comparison between Sentiment Accuracy CLMMs with and without confidence as a fixed effect

Based on these results, we can come to the following conclusions. We see that adding a second NMT output yields a large statistically significant improvement in relevance accuracy and a moderate though not statistically significant increase in analyst confidence, suggesting that Intervention 1 is effective at the relevance judgment level of comprehension and suitable for use in scanning tasks. Adding RBMT provides moderate statistically significant improvements in relevance and entity accuracy with minimal effect on analyst confidence, indicating some benefit at the relevance judgment level of comprehension and suggesting that perhaps RBMT provides particular benefit at the keyword-spotting level of comprehension needed for the entity element of our user study task. The combination of both a second NMT and RBMT provides a statistically significant benefit over a single NMT or NMT with RBMT for relevance judgment, but adding only the second NMT provides a greater benefit to relevance judgment and does not have a statistically significant effect on entity or sentiment accuracy. Analyst confidence appears to be well-calibrated with all three levels of accuracy, particularly sentiment accuracy. Although we do not see significant improvements in sentiment accuracy from either intervention or their combination, we do not see a miscalibration of confidence in sentiment accuracy from the interventions. We can infer that the interventions

do not provide enough improvement in deep comprehension to justify using them beyond the scanning use case, but they also do not lead to overconfidence in performance at this comprehension level, so implementing these interventions may not increase the temptation to engage in riskier MT use.

5.6.4 User Feedback

Participant responses to the pre- and post-task questionnaires provide further insight into the effects of adding a second NMT output and/or Motrans RBMT output. First, we discuss analyst trust, followed by display preferences.

Responses to key questions on the trust scale are summarized in Figures 5.8-5.13. Most participants agreed that the translations they typically see on the job and NMT1 translations were fairly consistent quality but found NMT2 and Motrans less consistent (Figure 5.8). Recall from Section 5.5.7.2 that NMT1 (NLLB) had the highest percentage of translations that were disconnected from the source text on the user study comments. NLLB's massively multilingual training data may enable it to produce more consistently fluent translations, even when the meaning is not correct. The high perception of consistency for translations analysts have seen on the job is likely due to having more exposure. Participants only saw 61 short translations from NMT2 and only 30 or 31 from Motrans, which may simply not be enough to get a feel for them.

On reliability (Figure 5.9), no individual MT system had higher than 50% agreement analysts could count on them to be correct, but both adding a second

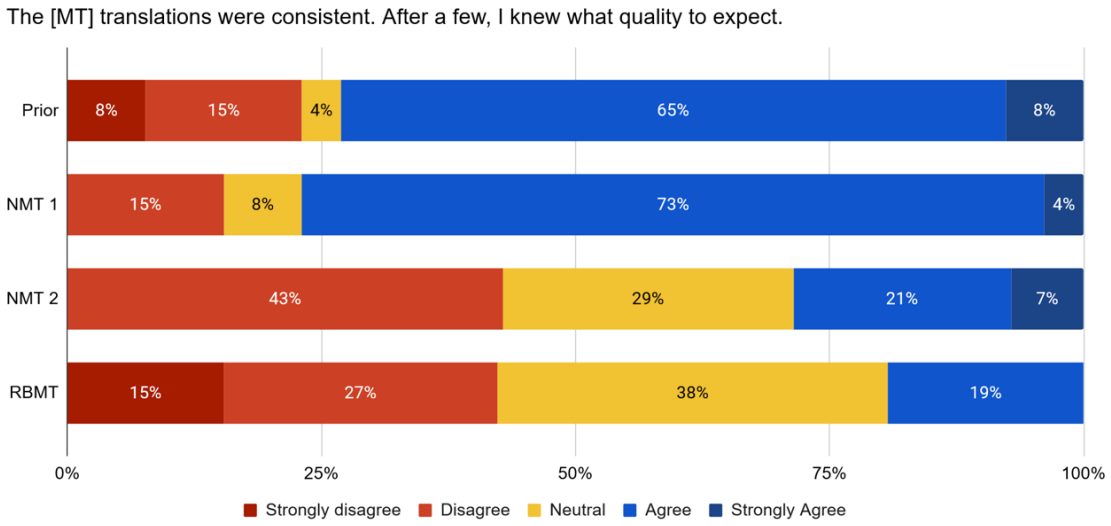


Figure 5.8: Participant responses to the consistency item.

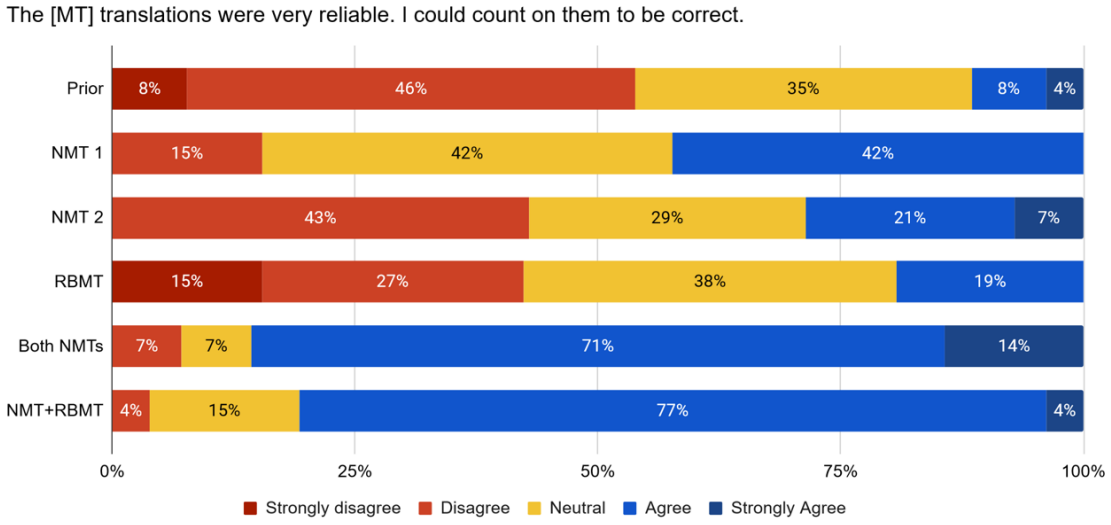


Figure 5.9: Participant responses to the reliability item.

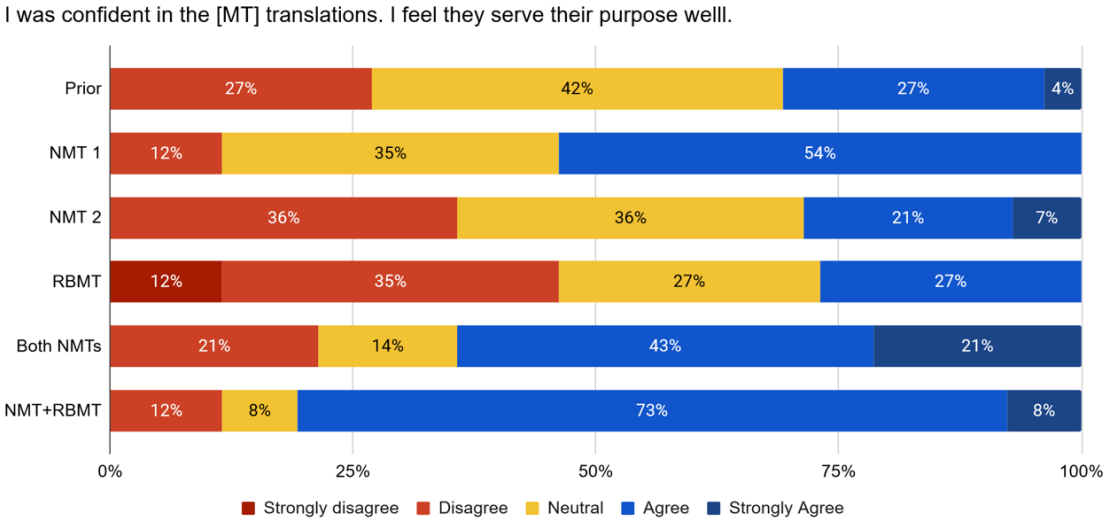


Figure 5.10: Participant responses to the confidence item.

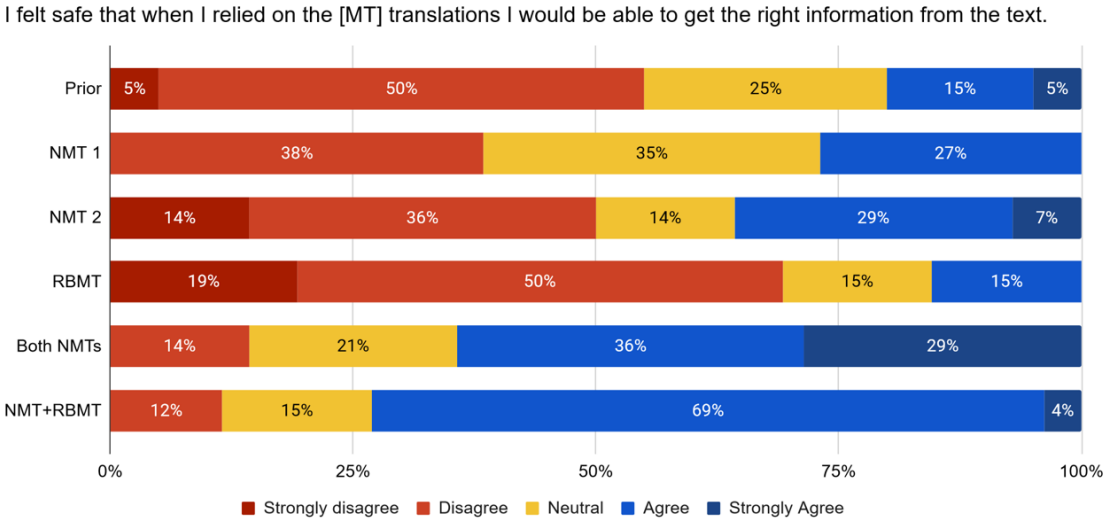


Figure 5.11: Participant responses to the safety item.

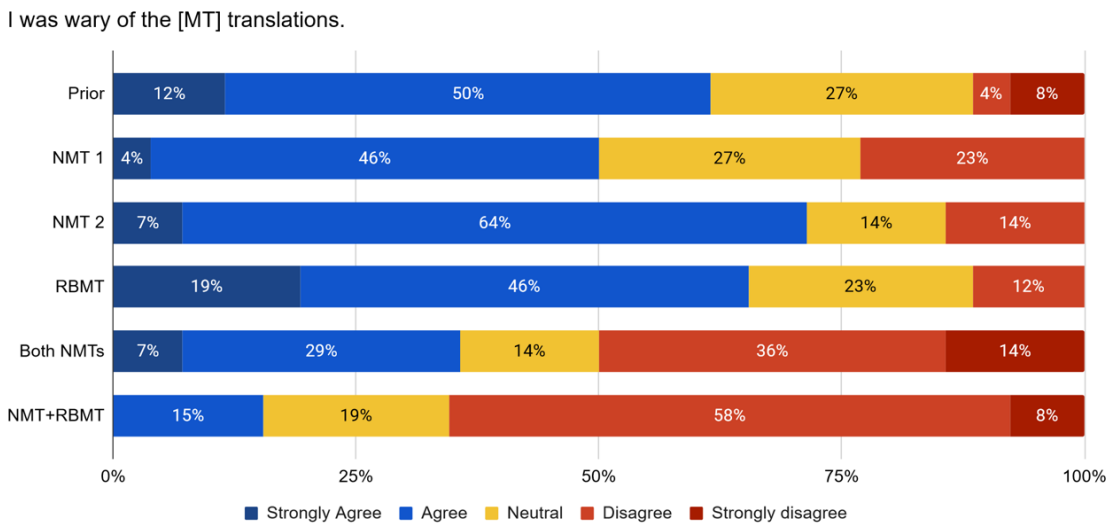


Figure 5.12: Participant responses to the wariness item.

NMT and adding RBMT had more than 80% agreement that they could count on at least one of the translations to be correct. This matches the high confidence scores we saw in the quantitative results. Figure 5.10 shows that NMT2 and RBMT both had just under 30% agreement that those systems alone served their purpose well, but 81% of participants agreed that translations from the combination of NMT and RBMT served their purpose well compared with 64% for the 2-NMT combination, which is not far above 54% for NMT1 alone. It may be that analysts see the RBMT output alone as not serving its purpose well for triage but as useful for keyword spotting when paired with NMT. Indeed, six analysts mentioned using Motrans this way in their open-ended comments. 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."

Figures 5.11 and 5.12 show how safe analysts felt when relying on the MT output and how wary they were of the output. The order of the scale in Figure

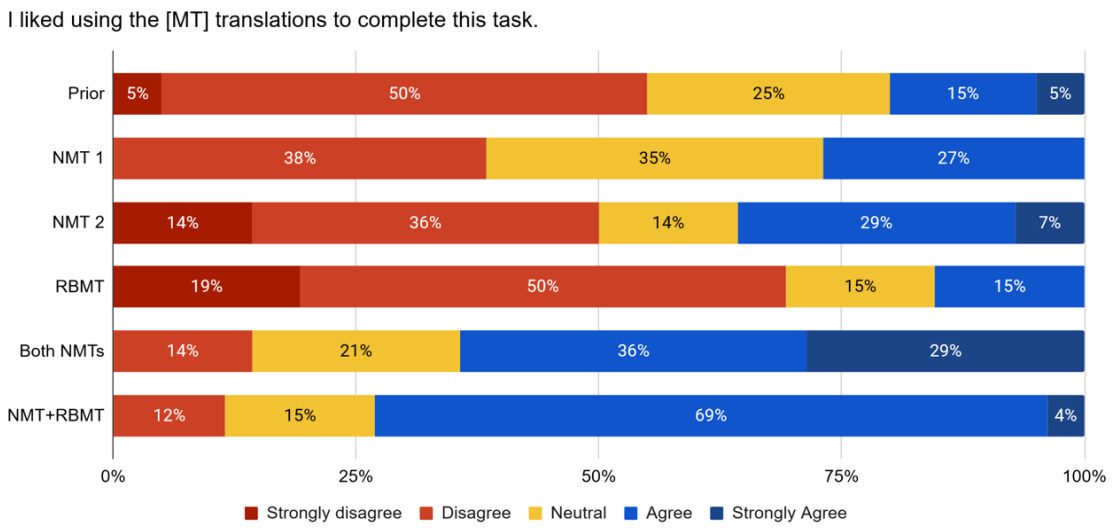


Figure 5.13: Participant responses to the likability item.

5.12 is reversed from the others because a participant agreeing that they felt safe is similar to disagreeing that they felt wary. Except for NMT2, we see that wariness and lack of safety track fairly closely. For NMT2, we see that 31% of participants agreed that they felt safe they would be able to get the right information from the text, but only 14% weren't wary of NMT2 translations. We also see in Figure 5.13 that analysts overwhelmingly liked the combination of NMT and RBMT (73%) but did not like using the RBMT output (69%). Likewise, they liked having two NMT outputs (65%) but did not like using the NMT2 output (50%). These seeming contradictions may be tied to how the analysts see themselves using the MT. The wariness statement refers only to the MT itself, but the safety statement puts it in the context of use. It may be that analysts feel safe that they can get the right information because they believe they are appropriately wary and will be able to evaluate the information effectively. Similarly, analysts seem to like having access to NMT2 or Motrans as long as they have something to compare against. Prior

work has indicated that intelligence analysts may be more likely than the general population to have an internal locus of control (Crouser et al., 2020), and that could explain the analysts' confidence that they will be able to take advantage of less-than-ideal MT output.

Finally, the survey asked participants whether they would like various ways of providing outputs from more than one MT system. Figure 5.14 shows that analysts in the 1-NMT setting would overwhelmingly like to see NMT and Motrans output side-by-side (83% agree, including 58% strongly agree). Analysts in the 2-NMT setting responded nearly identically to the ideas of two NMT outputs side-by-side and one NMT with Motrans, but there was slightly more enthusiasm for two NMT outputs and Motrans together (64% compared to 58%, a difference of only one analyst). Participants were not happy with the idea of a single MT output as default that could be replaced by a different MT output on demand, but they were less opposed if the system default was NMT (57%) rather than RBMT (76%) or if the user could choose the default (42%). The option that received the fewest votes of disagreement (one analyst) was allowing the user to choose more than one MT system to be side-by-side defaults. Together, these responses show strong support for providing outputs from multiple MT systems.

5.6.5 Time on Task

Before we conclude, we briefly address the issue of time on task. The user study platform tracked the median time on task for each user across all tasks, as

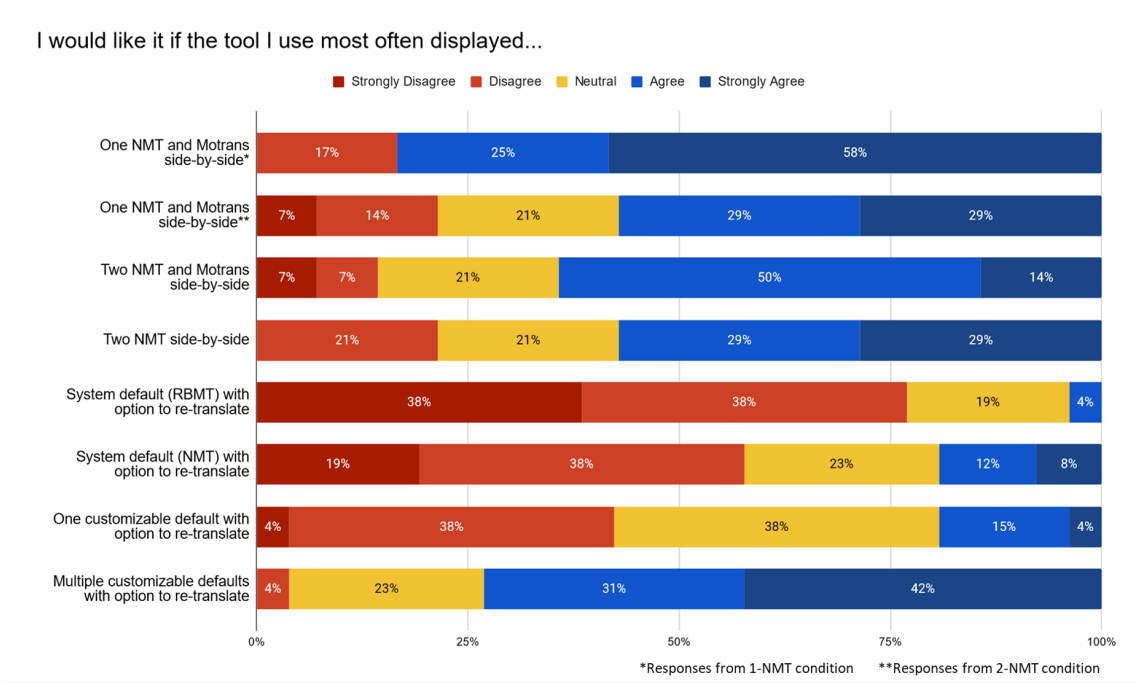


Figure 5.14: Participant responses to the display preference items.

shown in Figure 5.15. Because the times are not broken down by individual task, we cannot compare the with- and without-RBMT conditions, but we can compare 1-NMT and 2-NMT. We observe a slightly lower median time on task for the 1-NMT condition (13 minutes 37 seconds or about 54 seconds per comment) than for 2-NMT (14 minutes 31 seconds or about 58 seconds per comment) but a much larger range for the 1-NMT condition. The difference in medians is 54 seconds or approximately four additional seconds per comment to be labeled. This could reflect the additional time it takes to read a second NMT output, but the difference is not statistically significant, and confounding factors in the environment, such as interruptions or distractions, likely affected time on task. The similar times reported for 1-NMT and 2-NMT suggest that there is not *necessarily* a dramatic increase in median time required to incorporate information from a second NMT in relevance judgment

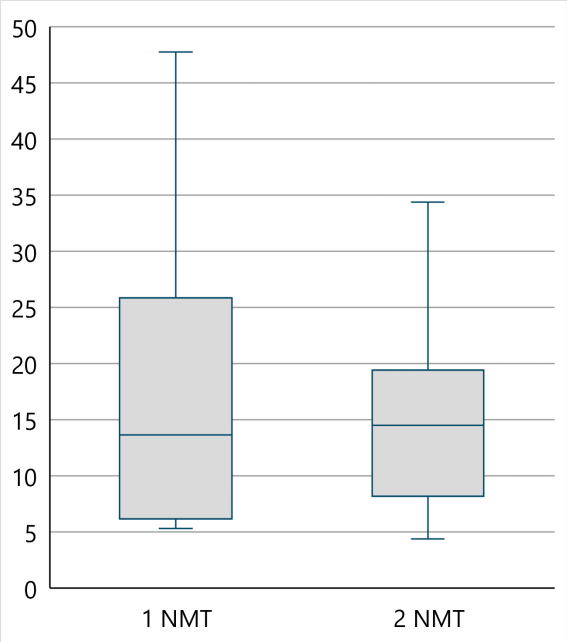


Figure 5.15: Time on task (minutes) for users in the 1-NMT setting compared to the 2-NMT setting.

decisions, but the actual time and effort required could vary widely depending on the length of the text and the way the MT outputs are displayed to the user. We will discuss this issue further in Section 5.8.

5.7 Conclusions

In this chapter, we introduced the MT-enabled scanning use case in the broader context of intelligence analysis and conducted a user study of two interventions designed to help analysts appropriately calibrate their trust in the MT output. 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. This level of comprehension is not typically needed for standard triage tasks but would be needed if analysts were to report off MT output.

Adding RBMT output has little effect on analyst confidence, but providing a second NMT output significantly improves it. This does not appear to be overconfidence, as confidence remains 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.

5.8 Limitations

This user study attempted to maintain realism to the greatest extent possible, but some of the aspects of that realism, as well as practical design decisions, limit the generalizability of the results.

In this study, participants were not timed and participated from their workstations during their work day. This is the same environment where they complete their typical MT-enabled scanning tasks, contributing to the realism of the study, but it prevents us from effectively measuring the impact of the interventions on the time it takes to complete those tasks.

The short length of texts in the user study allowed the participants to easily compare the side-by-side MT outputs and was similar to some MT-enabled scanning

use cases, but other use cases involve longer text. Additional user interface features may be required to help users recognize differences and similarities between the MT outputs. Most NMT models currently available translate on the sentence level, so displaying the outputs aligned by sentence may help users to compare, but newer models with larger context windows may be computationally difficult to align at that level. Features to help users identify similarities and differences between the outputs, such as those presented in Gero et al. (2024), could help users recognize potential errors in larger, unaligned MT outputs. This type of interface could also help users take advantage of more outputs than would be practical to read in a typical MT interface.

The user study was limited to a relatively small number of short translation examples from one language into English and used a small pool of intelligence analysts from a single U.S. government agency. The results can be seen as a starting point for considering ways to provide MT output to users with similar triage tasks, but the unique characteristics of these intelligence analysts and their workflows may limit the applicability of these results to other populations and use cases.

5.9 Recommendations

Based on the analyst preferences and the improvements in relevance judgment accuracy, we recommend that two MT outputs be displayed side-by-side wherever MT is provided to intelligence analyst users. RBMT, like 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. If one of those NMT systems provides an ability to dynamically include terminology lists at translation time, it is possible that it could replicate the entity recognition benefits of the RBMT system.

In light of the limited information from this user study related to the impact on longer document triage, we recommend that MT integrators conduct additional user testing 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.

Finally, 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.

Chapter 6: Conclusion

6.1 Summary

We began this work with the assumption that users of machine translation who do not understand the source text would be misled by fluently inadequate machine translation outputs. Under this assumption, we used automated fluency and adequacy metrics to compare the frequency of fluently inadequate translations in output from statistical and neural MT models and found that NMT was more prone to generating fluently inadequate output (Martindale et al., 2019). We then questioned the fluency assumption, focusing instead on believability as more directly applicable to whether or not MT users would be misled by inadequate output, and found that although grammatical fluency plays a significant role, other factors such as semantic plausibility in context affect believability and allow users to sometimes recognize fluently inadequate translations and not be misled (Martindale et al., 2021).

We then turned to a specific use case to address the impact of potentially misleading MT output and to test simple, practical interventions to improve user performance. We designed a user study with tasks analogous to typical foreign language text triage in intelligence analysis and recruited intelligence analysts from

the U.S. Intelligence Community to participate in the study. Analyst performance on relevance judgment improved significantly with the use of two NMT outputs rather than one and also improved with a rule-based MT output. Performance on relevant entity identification improved with the addition of rule-based MT, but neither RBMT nor a second NMT resulted in statistically significant improvement in comprehension at the sentiment level. This suggests that despite improvements in MT quality, errors remain and the mitigating effects of multiple MT outputs are not sufficient to enable use cases beyond the level of triage.

6.2 Implications and Future Work

In this final section, we address some of the implications of this work and opportunities for future work.

For those who seek to deploy Machine Translation to users who do not understand the source language, this work provides both encouraging and troubling insights. Neural Machine Translation can produce output that is disconnected from the input (Chapter 3), often referred to as "hallucinations". Because NMT models are uninterpretable black boxes, users may not anticipate these kinds of errors. However, as concerning as these outputs may be, they are often rendered unbelievable by being implausible in either local or global context (Chapter 4). Output from a second MT system (NMT or RBMT) can provide further context to validate or invalidate the potentially misleading outputs, as demonstrated in the use case of MT-enabled scanning for intelligence analysis (Chapter 5). Investigations

into these and other interventions for additional use cases could benefit the community by helping to clarify which interventions are most useful in which situations. Would longer, more technical documents require different interventions or adaptations of the proposed interventions such as different user interfaces? Would showing multiple outputs still benefit a population less motivated and with more external locus of control than intelligence analysts? These and other questions provide fertile directions for future user-centered MT research.

We also note that in the time since the experiments in Chapters 3 and 4, Large Language Models (LLMs) have overtaken much of the natural language processing world. These models demonstrate a remarkable ability to generate fluent text (Mahowald et al., 2024) and have shown promise in both producing (Jiao et al., 2023; Li et al., 2024) and evaluating (Kocmi and Federmann, 2023) translations. However, the larger model size, training data, and context window of LLMs all but guarantee more fluent and potentially more believable translations but still do not guarantee accurate translations.

As an example of a situation where an LLM produces fluent and perhaps even believable translation, Table 6.1 shows a prompt containing one of the Persian Farsi news comment threads from the user study in Chapter 5, a human translation of the thread, and some example translations generated by prompting the Mixtral 8x7B LLM (Jiang et al., 2024) to translate, resetting the context, and repeating the same translation prompt. Because LLMs are probabilistic rather than deterministic, the model can produce different outputs from the same input, as we see in these examples. The input text includes several non-standard spellings of non-Farsi

Prompt
User1: تنها راه نجات جهان کشتن پوتین آن هم به روش حلق آویز بصورت برهنه ومدت یک هفته در یکی از میادین کی اف هست
User2: معمولا اینایی که نظرات تند اینطوری میدن از اون کشته مرده های بایدن و پمپئو هستند
User3: همونطورکه تو یه جورخیالاتی هستی اونام یه جوردیگن. پوتین باید دردادگاه بین المللی محاکمه بشه حتما.فقط یک مشکل وجود داره که مجازات اعدام اونجا لغو شده
Please translate from Persian Farsi into English without explanation.
Human Tranlsation
User1: The only way to save the world is to kill Putin, and to do it by hanging him naked for a week in one of the squares in Kyiv.
User2: Usually the people who make harsh comments like that are devoted fans of Biden and Pompeo.
User3: Just like you are one kind of delusional, they're another kind. Putin should definitely be tried in the International Criminal Court. The only problem is that the death penalty has been abolished there.
Mixtral Translation A
User1: The only way to save the world is to kill Putin in a discreet way, for example by hanging him for a week in one of the MI6 buildings.
User2: Those who have such hardline opinions are usually either terrorists or from the NRA and pro-gun lobby.
User3: Just as you have fantasies like that, they have their own. Putin should definitely be put on international trial. There's just one problem, capital punishment has been abolished in those places.
Mixtral Translation B
User1: The only way to save the world is to kill Putin, even that could be through a slow and subtle hanging for a week in one of the Guantanamo Bay cells.
User2: Those who have such harsh views are usually just bloodthirsty and pumped up.
User3: Just as in some thoughts, they have some kind of phantasy. Putin must definitely be prosecuted in an international court. The only issue is that capital punishments have been abolished there.
Mixtral Translation C
User1: The only way to save the world is to kill Putin, slowly, in a sauna, for a week, in a kung fu style.
User2: Those who have such harsh views are usually hawks and warmongers.
User3: As you're imagining, they have some kind of grudge. Putin must definitely be put on trial in an international court. The only problem is that executions have been abolished there.

Table 6.1: An example Persian Farsi news comment thread, its human translation, and three independent translations produced by Mixtral 8x7B.

words including "کی اف" for Kyiv, "بایدن" for Biden, and "پمپئو" for Pompeo as well as Tehrani dialect phrases like "کشته مرده های" (devoted fans). Note that the prompt for translating contained all three messages in the thread, allowing the model to potentially recognize the context as an online comment thread. The model seems to have replaced the unknown entities with unrelated entities that were semantically plausible, given the context. The ability to fit the larger context may make these hallucinations more believable, though some disfluencies have also slipped in. Similar experiments to those in Chapters 3 and 4 could show whether inadequate LLM output like this is more often believable than inadequate NMT output.

The diversity of hallucinations from the same input text demonstrates the potential utility of approaches like Gero et al. (2024) for generating many outputs and indicating similarities and differences to the user. It is also worth noting that the specific prompt used can dramatically affect the output (Zhang et al., 2023), so observing stability in the output with changes in the prompt may also provide useful indicators of potential issues in the translation.

Other LLM-related interventions that could be investigated to mitigate misleading translation output include using LLMs to post-edit MT output (e.g., Xu et al., 2024; Raunak et al., 2023; Chen et al., 2024; Vidal et al., 2022; Ki and Carpuat, 2024) or using LLMs for quality estimation (e.g., Huang et al., 2024; Rei et al., 2023). Of particular interest is Fernandes et al. (2023), who propose an approach to using LLMs to label errors in MT output called AutoMQM. With the zero-shot prompt from Fernandes et al. (2023) and Mixtral Translation A from Table 6.1, OpenAI's GPT-4o successfully identified the major errors and most of the

minor errors, which would likely be sufficient to prevent a user from being misled. However, there remain hurdles to the widespread application of this approach. The success of the approach will likely depend on the quality of the underlying model and whether it has been exposed to enough text in the source and target languages in the relevant domain. These models can be costly whether measured in the computational resources required to deploy an open source LLM or in subscription fees and privacy risk from using commercial LLMs. All of these factors would need to be weighed against potential benefits for a specific use case.

6.3 Concluding Remarks

In this dissertation, we have sought to understand the prevalence, nature, and impact of potentially misleading output and whether a simple intervention can mitigate its effects on monolingual users. We find that even as overall MT quality has increased, the errors in MT output that do occur have become more fluent and often more believable, increasing the likelihood that users who do not understand the source language will be misled. Even simple mitigations like providing a second neural MT output or rule-based MT output can help users to more effectively perform triage tasks, but there is still room for improvement. We believe that the observations in this work can be applied to inform MT deployment decisions and carried forward as a foundation for further user-centered MT research with LLMs and whatever comes next in the evolution of machine translation.

Bibliography

ACTFL. 2012. ACTFL Proficiency Guidelines 2012.

Hirotogu Akaike. 1998. Information Theory and an Extension of the Maximum Likelihood Principle. In Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, New York, NY.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *In Proceedings the ACL, Ann Arbor*, volume 29, pages 65–72. Citeseer.

Yehoshua Bar-Hillel. 1951. The present state of research on mechanical translation. *American Documentation*, 2(4):229–237.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Yotam Berger. 2017. Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them'. *Haaretz*.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. 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.

Patrick Cadwell, Sharon O'Brien, and Carlos S. C. Teixeira. 2018. Resistance and accommodation: factors for the (non-) adoption of machine translation among pro-

fessional translators. *Perspectives*, 26(3):301–321. Publisher: Routledge _eprint: https://doi.org/10.1080/0907676X.2017.1337210.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Yue Cao, Hao-Ran Wei, Boxing Chen, and Xiaojun Wan. 2021. Continual Learning for Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3964–3974, Online. Association for Computational Linguistics.

Salvador Carrión and Francisco Casacuberta. 2022. Few-Shot Regularization to Tackle Catastrophic Forgetting in Multilingual Machine Translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 188–199, Orlando, USA. Association for Machine Translation in the Americas.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

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

Thomas Chesney and Daniel K. S. Su. 2010. The impact of anonymity on weblog credibility. *International Journal of Human-Computer Studies*, 68(10):710–718.

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*.

Anthony Cuthbertson. 2020. Facebook translates Chinese president's name to 'Mr Sh*thole'.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS ONE*, 14(9):e0222916.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven Word Alignment Interpretation for Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Kevin Duh. 2018. The Multitarget TED Talks Task.

Andrea Everard and Dennis F. Galleta. 2005. How Presentation Flaws Affect Perceived Site Quality, Trust, and Intention to Purchase from an Online Store. *Journal of Management Information Systems*, 22(3):56–95.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

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.

Mary Flanagan. 1994. Error classification for MT evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 65–72.

Kazimierz Florek, Jan Łukaszewicz, Julian Perkal, Hugo Steinhaus, and Stefan Zubrzycki. 1951. Sur la liaison et la division des points d'un ensemble fini. In *Colloquium mathematicum*, volume 2, pages 282–285.

B. J. Fogg. 2003. Prominence-interpretation theory: explaining how people assess credibility online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pages 722–723, Ft. Lauderdale, Florida, USA. Association for Computing Machinery.

B. J. Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, and Marissa Treinen. 2001. What makes Web sites credible? a report on a large quantitative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 61–68, New York, NY, USA. Association for Computing Machinery.

B. J. Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '99, pages 80–87, Pittsburgh, Pennsylvania, USA. Association for Computing Machinery.

M. L. Forcada, C. Scarton, L. Specia, B. Haddow, and A. Birch. 2018. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In *Proceedings of the Third Conference on Machine Translation*, volume 1, pages 192–203. ACL.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 449–458.

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. ArXiv:2401.13726 [cs].

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 33–41.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Yvette Graham, Christian Federmann, Maria Eskevich, and Barry Haddow. 2020a. Assessing Human-Parity in Machine Translation on the Segment Level. pages 4199–4207.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020b. Statistical Power and Translationese in Machine Translation Evaluation. pages 72–81.

Shuhao Gu and Yang Feng. 2020. Investigating Catastrophic Forgetting During Continual Training for Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. Technical report, Microsoft.

Donald K. Hawkins. 2017. Privacy Impact Assessment for the Refugee Case Processing and Security Vetting. Technical Report DHS/USCIS/PIA-068, U.S. Department of Homeland Security.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

David Hoffman. 2008. What happens when you lose everything.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects.

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:2401.06568 [cs].

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

United States Intelligence Community. 2015. Intelligence Cycle.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 624–635, New York, NY, USA. Association for Computing Machinery.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock,

Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. ArXiv:2401.04088 [cs].

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. ArXiv:2301.08745 [cs].

Blooma Mohan John, Alton Yeow-Kuan Chua, and Dion Hoe-Lian Goh. 2011. What Makes a High-Quality User-Generated Answer? *IEEE Internet Computing*, 15(1):66–71.

Douglas Jones, Martha Herzog, Hussny Ibrahim, Arvind Jairam, Wade Shen, Edward Gibson, and Michael Emonts. 2007. ILR-based MT comprehension test with multi-level questions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Prakruthi Karuna, Hemant Purohit, Özlem Uzuner, Sushil Jajodia, and Rajesh Ganesan. 2018. Enhancing Cohesion and Coherence of Fake Text to Improve Believability for Deceiving Cyber Attackers. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 31–40, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

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.

Soojung Kim. 2010. Questioners' credibility judgments of answers in a social question and answer site. *Information Research*, 15(2):15–2.

Soojung Kim and Sanghee Oh. 2009. Users' relevance criteria for evaluating answers in a social Q&A site. *Journal of the American Society for Information Science and Technology*, 60(4):716–727.

Margaret King, Andrei Popescu-Belis, and Eduard Hovy. 2003. FEMTI: creating and using a framework for MT evaluation. In *Proceedings of MT Summit IX, New Orleans, LA*, pages 224–231. Citeseer.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Katrin Kirchhoff, Daniel Capurro, and Anne M Turner. 2014. A conjoint analysis framework for evaluating user preferences in machine translation. *Machine translation*, 28(1):1–17.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Steven W. Knox. 2015. Extending pairwise element similarity to set similarity efficiently. *Presentation at MAA MathFest, Washington, DC*, 8.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-Informed Specificity and Semantic Plausibility for Dialogue Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3456–3466, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn. 2010. *Statistical machine translation*. Cambridge University Press, Cambridge ; New York.

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

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

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*.

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190. Association for Computational Linguistics.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2016. There Is No Logical Negation Here, But There Are Alternatives: Modeling Conversational Negation with Distributional Semantics. *Computational Linguistics*, 42(4):637–660.

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

Benjamin Lambert, Rita Singh, and Bhiksha Raj. 2010. Creating a linguistic plausibility dataset with non-expert annotators. In *Interspeech 2010*, pages 1906–1909. ISCA.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40(1):153–184.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

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 J. Liebling, Katherine Heller, Margaret Mitchell, Mark Díaz, Michal Lahav, Niloufar Salehi, Samantha Robertson, Samy Bengio, Timnit Gebru, and Wesley Deng. 2021. Three Directions for the Design of Human-Centered Machine Translation. Technical report.

Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet Needs and Opportunities for Mobile Translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Ziming Liu. 2004. Perceptions of credibility of scholarly information on the web. *Information Processing & Management*, 40(6):1027–1038.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació*, (12):455–463.

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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. pages 4791–4796.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 0(0).

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).

Mike McConnell. 2007. Intelligence Community Directive Number 203: Analytic Standards (Effective June 21, 2007). Technical Report ICD 203; Intelligence Community Directive 203, United States. Office of the Director of National Intelligence.

Peter McCullagh. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142.

Miriam J. Metzger, Andrew J. Flanagin, Keren Eyal, Daisy R. Lemus, and Robert M. Mccann. 2003. Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Annals of the International Communication Association*, 27(1):293–335.

Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. 2010. Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication*, 60(3):413–439.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain Robustness in Neural Machine Translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and

Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. ArXiv:2207.04672 [cs].

Mary Nurminen. 2019. Decision-making, Risk, and Gist Machine Translation in the Work of Patent Professionals. In *Proceedings of The 8th Workshop on Patent and Scientific Literature Translation*, pages 32–42.

Mary Nurminen and Niko Papula. 2018. Gist MT Users: A Snapshot of the Use and Users of One Online MT Tool. In *Annual Conference of the European Association for Machine Translation*, pages 199–208.

Douglas W. Oard and Philip Resnik. 1999. Support for interactive document selection in cross-language information retrieval. *Information Processing & Management*, 35(3):363–379.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2020a. Informative Manual Evaluation of Machine Translation Output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maja Popović. 2020b. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264, Online. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post, Yuan Cao, and Gaurav Kumar. 2015. Joshua 6: A phrase-based and hierarchical statistical machine translation system. *The Prague Bulletin of Mathematical Linguistics*, (104):5–16.

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.

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, Online. Association for Computational Linguistics.

Philip Resnik. 1997. Evaluating Multilingual Gisting of Web Pages. In *Proceedings of the AAAI Symposium on Natural Language Processing for the World Wide Web*, Stanford, CA. AAAI.

Soo Young Rieh and David R. Danielson. 2007. Credibility: A multidisciplinary framework. *Annual review of information science and technology*, 41(1):307–364.

Charles S. Robb, Laurence H Silberman, Richard C Levin, John McCain, Henry S Rowen, Walter B Slocombe, William O Studeman, Charles M Vest, Patricia Wald, and Lloyd Cutler. 2005. The Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction: Report to the President of the United States. Technical report, Commission on Intelligence Capabilities Regarding WMD, Washington, DC.

Laura Rossi and Dion Wiggins. 2013. Applicability and application of machine translation quality metrics in the patent field. *World Patent Information*, 35(2):115–125.

Carson Schütze. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Chenze Shao and Yang Feng. 2022. Overcoming Catastrophic Forgetting beyond Continual Learning: Balanced Training for Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Lin-*

guistics (Volume 1: Long Papers), pages 2023–2036, Dublin, Ireland. Association for Computational Linguistics.

Ross Smith. 2003. Overview of PwC/Sytranet on-line MT Facility. In *Proceedings of Translating and the Computer 25*, London, UK. Aslib.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, volume 200.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

SYSTRAN. 2021. Government Translation Solutions | SYSTRAN Technologies.

Irina P Temnikova. 2010. Cognitive Evaluation Approach for a Controlled Language Post–Editing Experiment. In *LREC*. Citeseer.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

John Tinsley. 2017. Machine Translation and the Challenge of Patents. In Mihai Lupu, Katja Mayer, Noriko Kando, and Anthony J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, The Information Retrieval Series, pages 409–431. Springer, Berlin, Heidelberg.

Antonio Toral. 2020. Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019. pages 185–194.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. pages 113–123.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Yeganeh Torbati. 2019. Google Says Google Translate Can't Replace Human Translators. Immigration Officials Have Used It to Vet Refugees.

Barak Turovsky. 2016. Ten years of Google Translate.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence measures for statistical machine translation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

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

David Vilar, Jia Xu, Luis Fernando d' Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *Proceedings of LREC*, pages 697–702.

Clare R. Voss and Calandra R. Tate. 2006. Task-based Evaluation of Machine Translation (MT) Engines. Measuring How Well People Extract Who, When, Where-Type Elements in MT Output. In *Proceedings of the 11th Annual conference of the European Association for Machine Translation*, Oslo, Norway. European Association for Machine Translation.

Peter de Vries, Cees Midden, and Don Bouwhuis. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, 58(6):719–735.

Chaojun Wang and Rico Sennrich. 2020. On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

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.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

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.

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.

Jin Yang and Elke D Lange. 1998. SYSTRAN on AltaVista a user study on real-time machine translation on the Internet. In *Conference of the Association for Machine Translation in the Americas*, pages 275–285. Springer.

X Jessie Yang, Christopher D Wickens, and Katja Hölttä-Otto. 2016. How users adjust trust in automation: Contrast effect and hindsight bias. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 196–200. SAGE Publications Sage CA: Los Angeles, CA.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study. In *Proceedings of the 40th International Conference on Machine Learning*, pages 41092–41110. PMLR.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).