Opportunities for Human-centered Evaluation of Machine Translation Systems

Daniel J. Liebling
Google Research
601 N. 34th St.
Seattle, WA 91803 USA
dliebling@google.com

Samantha Robertson[*]
University of California, Berkeley
253 Cory Hall
Berkeley, CA 94720 USA
samantha_robertson@berkeley.edu

Katherine Heller
Google Research
1600 Amphitheatre Parkway
Mountain View, CA 94043 USA
kheller@google.com

Wesley Hanwen Deng[†]
Carnegie-Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
hanwend@andrew.cmu.edu

Abstract

Machine translation models are embedded in larger user-facing systems. Although model evaluation has matured, evaluation at the systems level is still lacking. We review literature from both the translation studies and HCI communities about who uses machine translation and for what purposes. We emphasize an important difference in evaluating machine translation models versus the physical and cultural systems in which they are embedded. We then propose opportunities for improved measurement of user-facing translation systems. We pay particular attention to the need for design and evaluation to aid engendering trust and enhancing user agency in future machine translation systems.

1 Introduction

Evaluation of Machine Translation (MT) system performance seeks to close the gap between automated measurements and human judgments of translation quality. The MT evaluation literature contains a variety of work to quantify how closely human and automated measurement are linked. The models being evaluated in this literature are often complex combinations of traditional NLP and neural components. In this paper, we consider these systems as major but separate components of larger user-facing translation *systems* embedded in larger social contexts. These translation systems serve diverse user needs that go beyond simple translation of text to text. Grounded in research across domains, we motivate new research in three specific areas. First, translation system users, their motivations, and usage practices deserve further inquiry. Second, we argue that end-users (both individual and institutional) evaluate system performance differently from the system developers. End-users necessitate evaluation of the entire system, rather than just the model. Finally, we argue that significant research is necessary around nurturing trust in MT.

2 Machine translation users

Web- and app-based translation services, as well as translation APIs, are widely available. People who use these systems span geographic, socioeconomic, and occupational categories. We first discuss the contexts of consumers, a group of people who encounter machine translation as part of transacting their daily lives, but not as part of their vocation. Then we discuss the needs of translation professionals, people who use machine translation as part of their occupational workflows.

2.1 Primary consumers

We consider consumers as individuals who have specific needs for communicating across languages. As an additional factor, we consider individuals with "high-stakes" communication needs as a specific subset.

2.1.1 General public

Lack of language proficiency in a target language can be temporary (e.g. while traveling) or more persistent. In the former case, MT-based systems for travel translation have existed in various forms from personal digital assistants (PDAs) (Isotani

[*] Work while the author was an intern at Google Research.

[†] Work while the author was a student at University of California, Berkeley.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 229 - 241
July 10-15, 2022 ©2022 Association for Computational Linguistics

et al., 2002) to unidirectional offline speech-to-speech translation devices, to bidirectional devices embodied as earbuds that stream audio and translations to cloud services. Although these products view international travelers as their target market, the language needs of travelers are narrower than general purpose translation. Travelers' actual reported translation needs are lower than their perceived needs (Liebling et al., 2020).

Students use translation systems to complete homework assignments, especially for language-learning. Though educators hold varying opinions about this, translation systems are frequently adopted by second-language learners (Niño, 2009). The use of translation systems can improve students' lexical diversity (Fredholm, 2019) and the students have complex, nuanced perceptions of system quality (Xu, 2021).

Immigrants with language needs often turn to online or app-based translation when they have access to the associated technology (phones, laptops, etc) (Liebling et al., 2020). Proficient individuals also use machine translation. Using a second language for academic writing, which is domain-specific and highly structured, can be difficult; writers turn to machine translation to verify their word choices or even translate larger blocks of text (Bowker, 2020). Importantly, primary consumers may be aided by other individuals. For example, librarians will introduce the use of translation web sites or apps to immigrants with limited proficiency in their new country's language (Bowker and Buitrago Ciro, 2015). In this scenario, the first party (librarian) evaluates the quality of the system based on repeated interaction, whereas the second party (library patron) evaluates the quality based on a single interaction.

2.1.2 High-stakes use cases

A "high-stakes" scenario speaks to the potential impact and consequences of its outcomes. There is growing evidence that MT is widely used in high-stakes contexts such as healthcare (Vieira et al., 2020), law (Vieira et al., 2020; Kit and Wong, 2008), immigration (Torbati, 2019), and policing (Vieira et al., 2020; Liebling et al., 2020; Berger, 2017). Policymakers may discourage the use of MT () because the quality and reliability of machine translation in these settings is less well understood and the potential for harm from mistakes is high. As such, it is difficult to know how widespread these practices really are. For instance, many examples of MT use in policing have come to light only through journalism (Berger, 2017) and legal cases (Vieira et al., 2020).

The use of MT in the medical domain is comparatively better documented. Because language barriers in healthcare are associated with disparities in quality of care and health outcomes (Wilson et al., 2005), MT holds great potential for impact. When a language-concordant provider is not available, healthcare professionals are expected to use professional language services (Randhawa et al., 2013). For example, during covid-19 mass vaccination in 2021, the City of Seattle (United States) used a third-party live medical interpreting service provided over videoconference (Mattmiller, 2021). However, professional interpretation or translation can be difficult to arrange, especially in emergent situations, at smaller clinics, and for less commonly spoken languages (Turner et al., 2013; Şentürk et al., 2021). In these situations, healthcare providers have increasingly turned to MT (Turner et al., 2013; Randhawa et al., 2013). We return to perceived medical translation quality in Section 4.2.

2.2 Translation professionals

Translation professionals take content written in a source language and adapt it for use in a target language. They interact with this content on a transactional basis. This differentiates them from literary translators whose work is seen as a creative, artistic act (Benjamin, 1996; Hatim and Mason, 1990). In this section, we focus on the former.

Professional translators use software tools to manage their workflows. These computer-aided translation (CAT) systems aim to increase translation throughput. Specifically, translation memory (TM) features track previous translations of terms and phrases and suggest the prior translation, creating efficiency and consistency. Some tools augment the workflow with machine translation, as in "post-editing" workflows. In this workflow, source text is machine translated first, then corrected by professionals. Although post-editing can improve throughput and lower cost, human translators have resisted this task (O'Brien, 2017, 2012) as it changes the nature of the work from translator to editor, thereby reducing the agency of the worker. Complex issues of agency, quality, and workplace demands lead to multifaceted perception of machine translation.

3 Evaluating translation systems

Evaluating translation systems used by humans is difficult because the task, system design, and model performance are intertwined.

3.1 Measurement of translation systems

A common problem in systems measurement is disentangling whether effects are due to the manipulation of experimental variables, or due to complex interactions within the systems (both technical and human) (Olsen, 2007). For user-facing translation systems, it is important to distinguish between the performance of the underlying text-to-text translation *models* and the *systems* which they compose.

3.1.1 Model evaluation techniques

MT model performance on a corpus is frequently measured using the BLEU score (Papineni et al., 2002), amongst other metrics. Large increases in BLEU correlate with improved human judgment of translation quality. This makes it useful for driving progress in model architectures and training regimes. However, marginal gains in BLEU scores cease to be meaningful indicators of overall improvement; Callison-Burch et al. (2006) argued that improvements in BLEU were "neither necessary nor sufficient" for material improvements in model quality. Because of these limitations, other metrics exist, such as METEOR (Banerjee and Lavie, 2005) and COMET (Rei et al., 2020).

Techniques like quality estimation provide tools for predicting the quality of model output on a source string. Projects such as *OpenKiwi* (Kepler et al., 2019) provide interactive visualization of quality at the phrase and term level, but it remains unresolved what the effects are of showing quality estimates to end-users (Turchi et al., 2015). Recent papers provide evidence that showing quality metrics to end-users does not significantly impact perception or behavior (e.g. Turchi et al. (2015)). This argues that the metrics favored by developers aren't meaningful to end-users. Other subfields of applied artificial intelligence have come to similar conclusions. For example, Konstan and Riedl (2012) showed that the user experience in which recommender systems algorithms were embedded, significantly affected users' value perceptions.

3.2 Scale of studies

Similar to translation systems, web search engines are another example of how algorithms — in this case, ranking algorithms — are embedded in a larger system and context. Around the same time as the first web-based translation systems emerged, researchers published the first analysis of web search logs (Silverstein et al., 1999). Ranking metrics existed in the Information Retrieval (IR) community before web search engines (Salton and McGill, 1983), but the existence of search logs became a rich source of information on information-seeking behavior and use of search systems (Dumais et al., 2014; White, 2016). Various search log datasets were published, although widespread adoption of stricter privacy policies limited further releases. To our knowledge, there is no published research that uses machine translation logs to understand the behavior of online MT users. Instead, most research on how translation *systems* (versus *models*) comes from short-term user studies, ethnographic and qualitative investigation, and news reports. It seems unlikely that large-scale analyses will become available, so the burden of proof lies in small-scale studies.

3.3 Experimental protocols

Given that small-scale studies will likely compose the bulk of user-facing translation systems research, we now enumerate key challenges for reproducible systems research in this space. By systems evaluation, we mean studies where the MT model itself is fixed, but user interface affordances and interaction techniques are tested.

3.3.1 Participant selection

Second-language proficiency varies highly across individuals. In order to reduce experiment noise due to language ability, researchers need to find multilingual participants whose skills are roughly matched. A thorough experiment would also measure the proficiency of participants, but for simplicity, these skills are typically self-rated. Since model qualities vary across language pairs and direction of the translation, the ideal system setup will use language pairs with similar model quality.

Cultural backgrounds of those participants also vary. Wang et al. (2013) measured effects of machine translation on brainstorming between Chinese (Mandarin Chinese-speaking) and American (English-speaking) dyads. They measured quantity and frequency of idea generation in human- or MT-mediated dialogues. They noted that cultural bias towards or against quantity over quality might affect the results. The experimental design was asymmetrical; conversations were mediated through English in both conditions. Fully exploring the space of multilingual interaction requires evaluating systems across languages and linguistic backgrounds, which adds cost and time.

Pituxcoosuvarn and Ishida (2018) designed a system to allow multilingual users with varying degrees of proficiency to converse with each other through text. Participants could communicate through their L2 proficiency or through machine translation. In this way, the researchers could evaluate quality (measured via communication difficulty and verbosity) using mismatched participants.

3.3.2 Task design

Researchers can us a variety of tasks to encourage dialogue between participants, in order to study multi-party communication in translation. Of particular importance is eliciting spontaneous speech, as that is closer to how humans would use translation systems for communication in the real world. These needs are mismatched with the transactional affordances of contemporary translation web sites. Liebling et al. (2020) found that people used speech translation apps in a variety of rich, complex interactions from simple quotidian purchasing transactions to sensitive dialogues such as negotiating working conditions with an employer.

To elicit spontaneous speech, some established protocols exist. Pituxcoosuvarn et al. (2020) developed a MT-mediated chat app with a *desert-island survival* task (Lafferty, 1974). In this task, participants must choose and discuss items that they would desire should they be stranded on a desert island. The items are meant to elicit spontaneous responses from co-participants. Both Hara and Iqbal (2015) and Gao et al. (2014) used a *storytelling task* where participants were given a word list and encouraged to construct a story using those words. The *map task* (Anderson et al., 1991) gives participants paired maps; one participant's map contains a route and that person's goal is to get the other participant to follow the same route. Gao et al. (2015) used this task to study grounding in MT-mediated conversation.

These tasks give the experimenters some control over the conversation topic, which can reduce sources of error in across-subjects that comes from translation quality varying across different domains. If participants are given entirely open-ended tasks, their selection of domain will introduce additional noise. The tasks also have natural endpoints: all stimulus words are used, all differences are found, or route planning is complete. We argue that in the name of reproducibility, researchers should use existing tasks whenever possible, and share stimulus material resources as appropriate. Several of these conversation elicitation protocols have extensive validation in the social sciences. Because these tasks have been used for monolingual research as well, studies that employ them can serve as useful baselines when evaluating the effect of MT mediation on human conversations.

4 Trust in translation systems

Although some argue that MT has reached "human parity," many challenges remain. Output translations still contain errors, especially for lower resource languages. While minor errors in low-stakes contexts may be inconsequential, the unpredictability and lack of control over MT output raises concern regarding these systems' reliability in real world settings. Researchers and users have also documented systematic errors and biases in MT system output, which can result in longer term, representational harms to minoritized and oppressed groups (Stanovsky et al., 2019). In this section, we review important concerns in the design and use of safe, reliable, and useful machine translation systems. In the following section we identify open research questions at the intersection of HCI and MT that can begin to address some of these challenges.

Systematic errors and biases have been documented across NLP models and tasks. For example, Bolukbasi et al. demonstrated that word embeddings can learn stereotypical associations between gender and occupations (Bolukbasi et al., 2016). Machine translation models also struggle with semantic gender; for example, difficulties arise when translating a sentence from a language with gender-neutral pronouns to one with gendered pronouns while avoiding reinforcement of gender stereotypes (Stanovsky et al., 2019). In 2021, a major web translation site addressed this by allowing users to select from multiple target translations with different genders.[1] This gives the user agency to control the output of the system, compensating for issues with the underlying model.

[1] https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

4.1 Design asymmetries

Typically, machine translation models are constructed by people with access to substantial computation resources, embedded in systems with simple user interface wrappers. Early speech-to-speech translation devices had interfaces which mirrored standard web interfaces, consisting of two text boxes and language selection (Waibel et al., 2003; Zhou et al., 2003). These systems were evaluated mainly on performance of the speech and translation *models*, rather than system performance from a user's perspective. Notable exceptions include the *Verbmobil* project (Wahlster, 2000); designers identified many concerns that influenced system design, including recovery from error. Similarly, the *BBN TransTalk* device (Prasad et al., 2013), developed as part of the military-funded *TRANSTAC* research, also included a way for at least one party to estimate the quality of the translation. These systems are designed from a single perspective, and reflect the power dynamics in the environment (Risku et al., 2021; Paullada, 2020). People with low socioeconomic status use machine translation systems to access social capital (Liebling et al., 2020), but do not appear to be consulted during system design.

For professional translators, translation tools are often designed with throughput in mind, rather than the usability for end-users. Lagoudaki surveyed users of translation memory (TM) systems and concluded that "end-users' demands seem to have been only of subordinate interest" (Lagoudaki, 2006, 2008). Moorkens and O'Brien (2017) reached a similar conclusions through a separate survey and structured interview process.

Translation systems research will therefore benefit from direct user involvement, using techniques like participatory design (Muller and Kuhn, 1993; Schuler and Namioka, 1993) and value sensitive design (Friedman, 1996; Friedman et al., 2002). For example, we recruited 9 immigrants to the United States (4 men, 5 women) for a multi-day participatory design exercise. Participants first articulated their language journeys. Then, working in teams of three, they constructed fake translation devices out of foam and paper. The variety of devices created spoke to the need for translation devices designed for use in various environments and highlighted how design can reflect the asymmetry of translation needs across conversation participants.

4.2 Use in High-Stakes Settings

As outlined in section 2.1.2, MT is used when human interpreters are not available in high-stakes settings such as healthcare and policing. These settings pose unique challenges because of the power dynamics between parties, domain-specific language, and ramifications of precise word choice.

Several studies investigated whether MT is reliable enough to be used in healthcare settings. Findings from the provider's perspective are mixed; while some studies show promising performance for languages like Spanish and Chinese (Khoong et al., 2019), others found that MT was not considered reliable for lower-resource languages (Das et al., 2019; Taira et al., 2021). Beyond translation accuracy, providers expressed concern that entering sensitive healthcare-related information into a free translation website may violate patient privacy (Vieira et al., 2020).

Less work has examined patients' perspectives. Turner et al. simulated emergency scenarios with English-speaking healthcare workers and low-English proficiency (LEP) Spanish- or Chinese-speaking patients, communicating via either Google Translate or an app that provided pre-translated phrases (Turner et al., 2019). Patient-participants found it difficult and unpleasant to communicate their needs and understand instructions due to translation inaccuracies and delays in using the interface. Panayiotou et al. (2020) worked with both elder-care providers and older people with LEP to understand their perceptions of and attitudes towards using MT. While many participants were excited about the potential to improve their communication, they also raised concerns about translation accuracy and the difficulty of learning to use the technology.

Patient needs for translation are not limited to patient-provider interactions. People frequently turn to the web for healthcare information prior to, and soon after visits to healthcare providers (White and Horvitz, 2013). Some LEP immigrants selected providers based on availability of interpreter services, but the ancillary processes of accessing healthcare (such as calling to make an appointment) was an unmet need and opportunity for MT (Liebling et al., 2020).

Although there is consensus that professional translation or interpreters are always preferable in healthcare settings, there also seems to be acceptance that MT is a valuable last resort. Several studies, which included patient perspectives, recommend opting for tools that provide fixed, professionally translated phrases where possible, in order to guarantee accuracy and patient privacy (Spechbach et al., 2017; Turner et al., 2019; Panayiotou et al., 2019, 2020). When they choose to use MT, providers are urged to exercise caution. Some suggestions include using simple language (Khoong et al., 2019) and watching patients' facial expressions to detect confusion (Randhawa et al., 2013), although the effectiveness of these strategies has not been evaluated. This suggests a promising direction for future work is to develop systems that provide safeguards and guidance for generating high quality translations.

4.3 Methods for calibrating trust

Modern translation systems treat translation queries as a stateless, single-shot problem. While user insight can improve on this, popular interfaces do not allow users to introspect on system performance. The interfaces provide few, if any indicators of uncertainty or quality. Translation system providers frame support for language pairs as a binary encoding (supported or not) despite a wide distribution of quality across language pairs and directionality within a language pair (i.e. French → Japanese vs. Japanese → French).

Individuals understand that the outputs of translation systems are not perfect (Liebling et al., 2020), but strategies for overcoming system limitations vary. Bowker and Buitrago Ciro (2019) argue for *Machine Translation Literacy*, a set of strategies for understanding and working with the limitations of MT. In general, the lack of transparency with respect to system behavior makes it more difficult for end users to develop and validate such strategies. In light of this, some research has begun to explore how to surface more detailed information about system behavior to end-users.

Pituxcoosuvarn et al. (2020) developed a system to warn MT users if a term could cause cross-cultural misunderstanding. The warnings weakly influenced awareness of translation sensitivity. However, the warnings prompted some participants to change their word choice. A similar intervention was proposed by Miyabe and Yoshino (2011). They tested several visualizations of accuracy, but found that indicating this did not change the performance of the participants in terms of time to repair poor translations. However, participants anecdotally found some affordances more or less valuable. Picking the metric to visualize is important. Here, the visualization used a single value conflating accuracy and fluency. Research shows that fluency and adequacy are not equally considered (Martindale and Carpuat, 2018).

Moreover, users' perceptions need to be continually reassessed as underlying model architectures change. Literature from the Translation Studies discipline demonstrates that professional translators' impressions of MT are largely based on pre-neural models (O'Brien, 2012), but these impression still persist (Moorkens and O'Brien, 2017).

The interventions discussed so far focus on providing real-time feedback to end-users of MT systems. Complementing such approaches should be transparent information about overall model performance. Researchers have proposed various generic techniques for expressing model capabilities and limitations. For example, Mitchell et al. (2019) introduced *model cards*, intended to provide end-users and even indirect users with easily understandable information. One commercial computer vision API[2] shows examples of how low-quality lighting and visual occlusion can affect model performance.

Model-based performance metrics are not the only way to understand MT models and systems. Recent research questions the culpability of training and evaluation reference data as well (Freitag et al., 2020). Commonly-used corpora of reference translations can exhibit poor lexical and grammatical diversity (trending towards "translationese" (Graham et al., 2020)). Bommasani and Cardie (2020) demonstrated, in the context of summarization, how the evaluation of datasets can reveal the inconsistencies between the data's underlying properties and actual usages. One way to address this is to construct *Datasheets for Datasets* (Gebru et al., 2021). Inspired by standard datasheets for electronic components, Gebru et al. proposed documenting a dataset's motivation, composition, collection process, recommended uses, and so on.

We suggest that machine translation models can have similar model cards and datasheets. These would provide concrete examples of model and data boundaries while addressing folk-theories of model performance, which may or may not hold true. Additional research should evaluate the per-

[2] https://modelcards.withgoogle.com/object-detection

ception of these cards and the effects, if any, on how people select models and use machine translation systems. The best development of cards-like reporting may require tweaking cards in the MT setting, and options should be thoroughly investigated. In the spirit of translation, it's important to have these resources accessible in the same languages that the systems support. This will help address the perception that access to the best apps and content is only available through English-language interfaces (Karusala et al., 2018).

4.4 Users' mental models

HCI has a long tradition of using one's mental model of a system to understand one's experiences of the system (Norman, 2013). As AI systems have become common in consumer-facing products, HCI researchers have studied these mental models of various AI systems (Kocielnik et al., 2019; Liao et al., 2020; Khadpe et al., 2020; Wu et al., 2019). For example, Wu et al. (2019) studied YouTube content creators' mental models of YouTube algorithms that rank, filter, and recommend content. Through one-on-one interviews, they identified users' three main algorithmic characterizations of the algorithm, and how these characterizations could serve as a conceptual framework to create new interactions.

End-users have mental models of how the systems use the text of suggestions that translation professionals provide. For example, King (2019) attempted to offer corrections to Google Translate. Google's own blog post about Google Translate indicates that it "learns" from user suggestions (Turovsky, 2016). With those expectations set, King submitted various corrections through the web interface and then later re-evaluated the system output for the corrected strings. From the author's perspective, the system did not change.

5 Implications on Future HCI+MT Research

In this paper, we have examined groups of translation users and the role that MT plays in their lives. Common needs that cut across these user groups include reliability, trust, and agency, and we examine these in more detail here.

5.1 Trust

Use of MT in high-stakes settings underscores the importance of ensuring these systems are safe and reliable. Harms resulting from mistranslations range from inconvenience or embarrassment, to systematic representational harms. The most obvious way to reduce the frequency of such incidents is to improve the quality of the underlying translation model. Improvement in the accuracy of translation improves the user experience, and reduces the risk of harm. However, machine translation will never be perfect. We argue that improvement in the underlying models is neither sufficient for, nor necessarily indicative of, increasing the reliability and trust with these systems.

System safety engenders trust (Salem and Dautenhahn, 2015). Disentangling *safety* from model performance opens new questions for translation interface designers and human factors researchers. How can we design MT systems that are safer to use? An immediately feasible opportunity is to improve system transparency. Providing more information about system performance can help users calibrate their trust appropriately. Most MT systems remain opaque to end users, making it difficult for people to decide when they can rely on translation output.

At the model level, more detailed paradigms like Model Cards or Datasheets for Datasets could help resolve questions around how models perform, from where the data is sourced, and how systems incorporate user suggestions. These are effectively promises from translation services providers. Future research should investigate how users are influenced by and respond to these materials.

Trust is earned through repeated interaction. We argue that systems should provide more nuanced indications of uncertainty and quality. Those should be matched with user-facing affordances for acting on this information. For example, the user can see how alternate phrasing of their source text affects the output, but in a way that's comprehensible to someone with limited proficiency in the target language. This will require new designs and evaluation methods. For example, Zouhar et al. (2021) found that providing back-translations can improve user trust in outbound translations but did not correspond to an improvement in translation quality. As detailed in Section 3.3.2, careful task design will be necessary to prove the value of these affordances. Even if an affordance does not have a direct effect on language use in the system, it may still engender trust. Unfortunately, the direct effects of trust are difficult to measure in laboratory studies; similar mixed effects are seen in human-robot interaction (Flook et al., 2019). Research on trust in machine translation should focus on specific user sets (e.g. students, immigrants, translation professionals) and circumstances.

Reliability issues are encountered when the machine translation system does not give expected results, with potentially negative consequences for the user. Understanding the limitations of machine translation and knowing when not to rely on the automated results are key to safe machine translation usage. Some of the "algorithmic disillusionment" (Eslami et al., 2018) users may experience with machine translation systems can also engender critical and vigilant usage of machine translation tools, especially in high-stakes scenarios.

5.2 Agency

We believe that carrying a sense of agency through the translation process will be a key part of future machine translation systems. We envision a world where individuals can effectively communicate across language boundaries without having to adapt their language use to suit the limitations of the system (e.g. by the use of "controlled language"). Although an utterance in one language may not have a precise equivalent in the target language, systems should provide affordances to help users understand those constraints. This will involve the design, implementation, and evaluation of new user experiences in translation systems. For example, current speech-to-speech translation systems assume that speech recognition is high-quality, and pass the transcripts directly to translation models. However, even human interpreters will ask the original speaker clarifying questions to best formulate a translation. Some MT models can incorporate longer context, but the translation UIs do not yet support ways to input context or clarify ambiguities in content or register. All these possibilities create large potential for human factors-informed research in the translation space.

For translators who intend to augment their work with machine translation technology, new interactions and interfaces are needed to establish an effective human-AI collaboration relationship. For example, Interactive Neural Machine Translation (Santy et al., 2019) provides human translators with on-the-fly hints and suggestions. However, recent research in human-AI collaboration showed that, when paired up with AI assistant, people might tend to over-rely on the AI system (Buçinca et al., 2021; Bansal et al., 2021). How to maintain the agency of human translators while effectively making suggestions remains an open question in this line of research.

5.3 Limitations

Our focus in this paper is on enumerating the contributions which HCI research can offer for safety and user-focused design, but it is equally important to recognize the limitations of what is achievable with improved design processes, user interfaces, or evaluations. In conclusion, we discuss limitations, highlighting complementary alternatives.

As we consider improving systems for use in high-stakes settings, we must also recognize fundamental concerns around the use and development of translation technology. Assuming that translation technology continues to improve, how will high-stakes users of machine translation understand the limitations? Under what conditions would machine translation be acceptable in medicine, policing, or immigration? How can we ensure the consent and center the needs of disempowered users in these interactions? Ethnographic and qualitative work with impacted communities, such as transnational migrants with low-language proficiency, can help establish consistent boundaries and good practices. Researchers must also consider how power dynamics shape the use and consequences of MT and avoid overly narrow conceptions of their users and those who are impacted.

6 Conclusion

Human communication is subtle and complex. For high-resource languages, machine translation models can successfully map many phrases across languages with high fluency and adequacy. We believe that the next frontier of research in machine translation is, through a human factors lens, to bring modern themes of trust and agency to the table. First, more robust qualitative and quantitative research is necessary to understand how people *interact* with MT systems today, and what improves the quality of those interactions. Machine translation users vary from professionals incidentally using MT as part of their workflow, to immigrants with low language proficiency who rely on MT systems to survive. From a design and implementation standpoint, we acknowledge that MT user interfaces have evolved very little. Given that people use machine translation to communicate in high-stakes scenarios, there is a great opportunity to provide new affordances which ultimately engender trust and facilitate more authentic, trustworthy communication, and avoid algorithmic harm.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Walter Benjamin. 1996. *The Task of the Translator*. Belknap Press of Harvard University Press, Cambridge, Mass., USA.

Yotam Berger. 2017. Israel arrests palestinian because facebook translated 'good morning' to 'attack them'. *Haaretz*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096.

Lynne Bowker. 2020. Chinese speakers' use of machine translation as an aid for scholarly writing in english: a review of the literature and a report on a pilot workshop on machine translation literacy. *Asia Pacific Translation and Intercultural Studies*, 7(3):288–298.

Lynne Bowker and Jairo Buitrago Ciro. 2015. Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies*, 10(2):165–186.

Lynne Bowker and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Prithwijit Das, Anna Kuznetsova, Meng'ou Zhu, and Ruth Milanaik. 2019. Dangers of machine translation: The need for professionally translated anticipatory guidance resources for limited english proficiency caregivers. *Clinical Pediatrics*, 58(2):247–249.

Susan Dumais, Robin Jeffries, Daniel M. Russell, Diane Tang, and Jaime Teevan. 2014. *Understanding User Behavior Through Log Data and Analysis*, pages 349–372. Springer New York, New York, NY.

Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA. Association for Computing Machinery.

Rebecca Flook, Anas Shrinah, Luc Wijnen, Kerstin Eder, Chris Melhuish, and Séverin Lemaignan. 2019. On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy? *Interaction Studies*, 20(3):455–486.

Kent Fredholm. 2019. Effects of Google Translate on lexical diversity: vocabulary development among learners of Spanish as a foreign language. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 13(26):98–117.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Batya Friedman. 1996. Value-sensitive design. *Interactions*, 3(6):16–23.

Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods.

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, page 852–863, New York, NY, USA. Association for Computing Machinery.

Ge Gao, Naomi Yamashita, Ari MJ Hautasaari, Andy Echenique, and Susan R. Fussell. 2014. Effects of public vs. private automated transcripts on multiparty communication between native and non-native english speakers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 843–852, New York, NY, USA. Association for Computing Machinery.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM (CACM)*, 64(12):86–92.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81. Association for Computational Linguistics.

Kotaro Hara and Shamsi T. Iqbal. 2015. Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3473–3482, Seoul, Republic of Korea. Association for Computing Machinery.

Basil Hatim and Ian Mason. 1990. *Discourse and the Translator*. Routledge.

Ryosuke Isotani, Kiyoshi Yamabana, Shin-ichi Ando, Ken Hanazawa, S. Ishikawa, Tadashi Emori, Kenichi Iso, Hiroaki Hattori, Akitoshi Okumura, and Takao Watanabe. 2002. An automatic speech translation system on pdas for travel conversation. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 211–216.

Naveena Karusala, Aditya Vishwanath, Aditya Vashistha, Sunita Kumar, and Neha Kumar. 2018. "Only if you use English you will get to more things": Using smartphones to navigate multilingualism. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26.

E.C. Khoong, E. Steinbrook, C. Brown, and A. Fernandez. 2019. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Internal Medicine*, 179(4):580–582.

Katherine M. King. 2019. Can Google Translate be taught to translate literature? A case for humanists to collaborate in the future of machine translation. *Translation Review*, 105(1):76–92.

Chunyu Kit and Tak Ming Wong. 2008. Comparative evaluation of online machine translation systems with legal texts. *Law Library Journal*, 100:299–321.

Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1):101–123.

J. Clayton Lafferty. 1974. *The Desert Survival Situation Problem: a Group Decision Making Experience for Examining and Increasing Individual and Team Effectiveness*. Human Synergistics.

Elina Lagoudaki. 2006. Translation memories survey 2006: Users' perceptions around TM use. In *Proceedings of ASLIB Translating and the Computer*, volume 28.

Elina Lagoudaki. 2008. *Expanding the Possibilities of Translation Memory Systems: From the Translator's Wishlist to the Developer's Design*. Ph.D. thesis, Imperial College, London.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

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, New York, NY, USA. Association for Computing Machinery.

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.

Michael Mattmiller. 2021. Personal communication.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Mai Miyabe and Takashi Yoshino. 2011. Can indicating translation accuracy encourage people to rectify inaccurate translations? In *Human-Computer Interaction. Interaction Techniques and Environments*, pages 368–377, Berlin, Heidelberg. Springer Berlin Heidelberg.

Joss Moorkens and Sharon O'Brien. 2017. Assessing user interface needs of post-editors of machine translation. In Dorothy Kenny, editor, *Human Issues in Translation Technology*. Routledge.

Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

Ana Niño. 2009. Machine translation in foreign language learning: language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL*, 21(2):241–258.

Donald A. Norman. 2013. *The design of everyday things*. Basic Books.

Sharon O'Brien. 2012. Translation as human–computer interaction. *Translation Spaces*, 1:101–122.

Sharon O'Brien. 2017. Machine translation and cognition. In John W. Schwieter and Aline Ferreira, editors, *The Handbook of Translation and Cognition*, chapter 17, pages 313–331. Wiley, Hoboken, NJ, USA.

Dan R. Olsen. 2007. Evaluating user interface systems research. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07, pages 251–258, New York, NY, USA. Association for Computing Machinery.

A. Panayiotou, A. Gardner, S. Williams, E. Zucchi, M. Mascitti-Meuter, A.M. Goh, E. You, T.W. Chong, D. Logiudice, X. Lin, B. Haralambou, and F. Batchelor. 2019. Language translation apps in health care settings: Expert opinion. *JMIR mHealth uHealth*, (4).

A. Panayiotou, K. Hwang, S. Williams, T.W.H. Chong, D. LoGiudice, B. Haralambous, X. Lin, E. Zucchi, M. Mascitti-Meuter, A.M.Y. Goh, E. You, and F. Batchelor. 2020. The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *Journal of Clinical Nursing*, 29(17-18):3516–3526.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amandalynne Paullada. 2020. How does machine translation shift power? In *Resistance AI Workshop at NeurIPS 2020*.

Mondheera Pituxcoosuvarn and Toru Ishida. 2018. Multilingual communication via best-balanced machine translation. *New Generation Computing*, 36(4):349–364.

Mondheera Pituxcoosuvarn, Yohei Murakami, Donghui Lin, and Toru Ishida. 2020. Effect of cultural misunderstanding warning in MT-mediated communication. In *Collaboration Technologies and Social Computing*, pages 112–127. Springer International Publishing.

Rohit Prasad, Prem Natarajan, David Stallard, Shirin Saleem, Shankar Ananthakrishnan, Stavros Tsakalidis, Chia lin Kao, Fred Choi, Ralf Meermeier, Mark Rawls, Jacob Devlin, Kriste Krstovski, and Aaron Challenner. 2013. BBN TransTalk: Robust multilingual two-way speech-to-speech translation for mobile platforms. *Computer Speech & Language*, 27(2):475–491. Special Issue on Speech-Speech translation.

Gurdeeshpal Randhawa, Mariella Ferreyra, Rukhsana Ahmed Omar Ezzat, and Kevin Pottie. 2013. Using machine translation in clinical practice. *Canadian Family Physician*, 59(4).

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.

Hanna Risku, Jelena Milošević, and Regina Rogl. 2021. *Responsibility, powerlessness, and conflict: An ethnographic case study of boundary management in translation*, volume 157 of *Benjamins Translation Library*, pages 145–168.

Maha Salem and Kerstin Dautenhahn. 2015. Evaluating trust and safety in HRI: Practical issues and ethical challenges. In *The Emerging Policy and Ethics of Human Robot Interaction*.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive neural machine translation prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 103–108, Hong Kong, China. Association for Computational Linguistics.

Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

Emre Șentürk, Mukadder Orhan-Sungur, and Tülay Özkan Seyhan. 2021. Google Translate: Can it be a solution for language barrier in neuraxial anaesthesia? *Turkish Journal of Anaesthesiology & Reanimation*, 49(2):181–182.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12.

Hervé Spechbach, Ismahene Sonia Halimi Mallem, Johanna Gerlach, Nikolaos Tsourakis, and Pierrette Bouillon. 2017. Comparison of the quality of two speech translators in emergency settings : A case study with standardized arabic speaking patients with abdominal pain. *European Congress of Emergency Medicine, (EUSEM 2017)*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Breena R. Taira, Vanessa Kreger, Aristides Orue, and Lisa C. Diamond. 2021. A pragmatic assessment of Google Translate for emergency department instructions. *Journal of General Internal Medicine*.

Yeganeh Torbati. 2019. Google says Google Translate can't replace human translators. immigration officials have used it to vet refugees. *Pro Publica*.

Marco Turchi, Matteo Negri, and Marcello Federico. 2015. MT quality estimation for computer-assisted translation: Does it really help? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 530–535, Beijing, China. Association for Computational Linguistics.

Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. *JMIR Public Health and Surveillance*, 5(1).

Anne M. Turner, Hannah Mandel, and Daniel Capurro. 2013. Local health department translation processes: potential of machine translation technologies to help meet needs. *AMIA Annual Symposium Proceedings*, 2013:1378–1385.

Barak Turovsky. 2016. Ten years of Google Translate.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2020. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, pages 1–18.

Wolfgang Wahlster. 2000. *Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System*, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg.

Alex Waibel, Ahmed Badran, Alan W. Black, Robert Frederking, Donna Gates, Alon Lavie, Lori Levin, Kevin Lenzo, Laura Mayfield Tomokiyo, Juergen Reichert, Tanja Schultz, Dorcas Wallace, Monika Woszczyna, and Jing Zhang. 2003. Speechalator: Two-way speech-to-speech translation in your hand. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Demonstrations*, pages 29–30.

Hao-Chuan Wang, Susan Fussell, and Dan Cosley. 2013. *Machine Translation vs. Common Language: Effects on Idea Exchange in Cross-Lingual Groups*, page 935–944. Association for Computing Machinery, New York, NY, USA.

Ryen W. White. 2016. *Interactions with search systems*. Cambridge University Press.

Ryen W. White and Eric Horvitz. 2013. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *Journal of the American Medical Informatics Association*, 21(1):49–55.

Elisabeth Wilson, Alice Hm Chen, Kevin Grumbach, Frances Wang, and Alicia Fernandez. 2005. Effects of limited English proficiency and physician language on health care comprehension. *Journal of General Internal Medicine*, 20(9):800–806.

Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, gatekeeper, drug dealer: How content creators craft algorithmic personas. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Jun Xu. 2021. Google Translate for writing in a Japanese class: What students do and think. *Journal of the National Council of Less Commonly Taught Languages (JNCOLTL)*, 30:136–182.

Bowen Zhou, Yuqing Gao, Jeffrey Scott Sorensen, D. Déchelotte, and M. Picheny. 2003. A hand-held speech-to-speech translation system. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 664–669.

Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021. Backtranslation feedback improves user confidence in MT, not quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 151–161, Online. Association for Computational Linguistics.