A Research-Based Guide for the Creation and Deployment of a Low-Resource Machine Translation System

John E. Ortega
Northeastern University
Boston, MA, 02115
USA
j.ortega@northeastern.edu

Kenneth W. Church
Northeastern University
Boston, MA, 02115
USA
k.church@northeastern.edu

Abstract

The machine translation (MT) field seems to focus heavily on English and other high-resource languages. Though, low-resource MT (LRMT) is receiving more attention than in the past. Successful LRMT systems (LRMTS) should make a compelling business case in terms of demand, cost and quality in order to be viable for end users. When used by communities where low-resource languages are spoken, LRMT quality should not only be determined by the use of traditional metrics like BLEU, but it should also take into account other factors in order to be inclusive and not risk overall rejection by the community. MT systems based on neural methods tend to perform better with high volumes of training data, but they may be unrealistic and even harmful for LRMT. It is obvious that for research purposes, the development and creation of LRMTS is necessary. However, in this article, we argue that two main workarounds could be considered by companies that are considering *deployment* of LRMTS in the wild: human-in-the-loop and sub-domains.

1 Introduction

This research-based guide surveys the literature in order to provide a guide for companies that plan on deploying low-resource machine translation systems (LRMTS) in the wild. The guide is meant to be used as a practical manner of knowing whether or not the LRMTS meets the minimum requirements established by the literature to support those who live in regions where the respective low-resource language is spoken. Much of the work in computational linguistics and machine translation (MT) focuses on high-resource languages, and especially English. In a recent ACL-2022 conference (Muresan et al., 2022) and MT workshop (WMT-2022 (Koehn et al., 2022)), there is considerable interest in "the Bender rule" (Bender et al., 2021) which states that the research community should move beyond English and even beyond high-resource languages. There are a number of commercial MT products that support an amazingly large set of language combinations, and there are some research groups that are attempting to support even more combinations (Costa-jussà et al., 2022). Of course, some language pairs are more successful than others. Some of the low-resource language pairs that end up being deployed in the wild can be considered useful and others not so useful or even downright unethical (Mager et al., 2023; Joshi et al., 2019) due to their low quality.

High-quality MT systems are more often than not back by neural networks; thus, neural machine translation (NMT) has advanced the state-of-the-art (SOTA) on many benchmarks. This is particularly true for high-resource languages like English and Spanish because neural methods have been shown to work best with huge amounts of data (Koehn and Knowles, 2017). More traditional methods such as phrase-based statistical machine translation (SMT) tend to work better than NMT when training data is limited. In this article we first explore in Section 2 a list of challenges for companies that are considering deploying a LRMTS in the wild. Secondly, we discuss in Section 3 the minimal requirements that a company should take into consideration when deploying an LRMTS. After presenting the challenges and minimum requirements, we provide an overview of related work in Section 4 to provide insight into the quality standards in Section 5 and how to address them in Section 6.

2 Challenge List

We argue that, despite a popular opinion that deploying LRMTS quickly is necessary for success (Bali et al., 2019), companies that deploy LRMTS should consider reviewing literature such as this article to address ethical and responsible concerns

Proceedings of Recent Advances in Natural Language Processing, pages 813–823
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_088

in order to avoid outright rejection by the low-resource community that their system targets. From the company's perspective, successful LRMTS require a compelling business case in terms of demand, cost and quality. Companies are more likely to fund projects that address those concerns. But, since quality tends to increase with the size of the training set (Koehn and Knowles, 2017) in NMT and even SMT, it can be hard to determine whether or not a LRMTS should be deployed in the wild. To avoid rejection of a LRMTS's deployment from its targeted community, we propose two workarounds: (a) human-in-the-loop and (b) subdomains to address the following three challenges that a LRMTS's creator *must* overcome as a first (not only) step:

> **Challenge 1.** The business case needs to be compelling in terms of demand, cost, and quality.
> **Challenge 2.** The LRMTS's quality should be good enough to provide value to its target community.
> **Challenge 3.** Workarounds should be considered when MT quality is low.

3 Minimum Viable Product (MVP): Minimal Requirements

While high-resource languages can be considered more reliable for MT, most LRMTS are probably not up to par for deployment in their respective target communities. We argue that LRMTS deployed for the wrong reason may cause more harm than help. If the needs of the of the low-resource community are not taken into account, results can be disastrous and difficult to turn around (Haroutunian, 2022). At a minimum, the questions and statements below should be addressed.

What if the low-resource community is not interested? Risks associated with widespread adoption of digital system deployed in the wild, such as Risks 1.0 and 2.0 defined by (Church et al., 2022), can be costly. It is a mistake to deploy LRMTS into the wild without sufficient demand. The MVP requires hundreds (if not thousands) of users in the low-resource community that are willing to use it. The ethical concerns could by far be more important than any other factor (Mager et al., 2023). When a company creates a business case for deploying a LRMTS, it should at a minimum take the following into consideration: (1) demand (market size), (2) costs (memory footprint and computation) and (3) high quality translations for ethical reasons.

Estimates of Demand. Demand for LRMTS seems to be low due to the lack of funding from nations where low-resource languages are spoken. While there are exceptions such as the European low-resource projects Horizon[1] and others, smaller countries with less governmental power like Peru, for example, provide less funding in general. (Camacho and Zevallos, 2020) Demand is focused on high-resource languages which have more speakers with more buying power. Nonetheless, with the introduction of large-language models (LLMs), interest by larger private companies like Meta (Costa-jussà et al., 2022) in LRMT has increased.

Business demand, while not easily calculable for low-resource regions, can occur in unforeseen situations. Crises situations, such as natural disasters, could constitute enough demand but much harder to forecast. (Cadwell, 2021) Unfortunately, these types of disasters can produce a higher demand in regions where low-resource languages are spoken and should be considered of upmost importance.

Estimates of Costs. Costs depend on many factors including computing resources. Due to the lack of data, LRMTS often attempt to leverage large-language models (LLMs) for additional performance but LLMs may be too expensive for practical deployments (Diddee et al., 2022). In addition to costs, LLMs introduce some more concerns (Marcus and Davis, 2020). Human-in-the-loop techniques can address some of these concerns, though such techniques tend to increase costs.

Estimates of Quality. Quality tends to increase with the size of the training set. How many parallel sentences are considered low-resource? We suggest these rules of thumb as a loose guide but company's should research more:

- low resource: ≈ under 300k (Weller-di Marco and Fraser, 2022; Tars et al., 2022)
- medium resource: ≈ 300k – 3M (Ortega et al., 2022)
- high resource: over 3M (Jonsson et al., 2020)

Variability of sentence length is an addition consideration that can cause trouble when systems are deployed in the wild. For low-resource languages,

[1] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en

it is often not feasible to improve quality by increasing the size of the training set. Section 5 will suggest two workarounds: (a) human-in-the-loop and (b) subdomains. As will be discussed in Section 5, quality not only includes standard metrics such as BLEU (Papineni et al., 2002), but also other considerations that may be more difficult to quantify such as biases and respect for cultural diversity.

Human informants can improve quality in a couple of ways. A LRMTS *must* have annotators to provide feedback on the quality of translations before deploying a system. Similar to others (Castilho et al., 2018; Way, 2018), the quality must be assessed and agreed upon before delivery. Sometimes, as described in seminal work by (Läubli et al., 2018), work can be crowdsourced. Whether the LRMTS be evaluated by crowdsourced humans or experts in linguistics or a few native speakers as was done in the work by (Ortega et al., 2020), any LRMTS that is to be deployed should include at a minimum well-versed annotator in the LRMTS. In addition, humans can reject inevitable bad outputs, as suggested by (Ebrahimi et al., 2023).

4 Related Work

This article is inspired by (Koehn and Knowles, 2017). Their work was written at a time when neural-based systems were catching up to statistical alternatives, but their paper helped to close the gap by identifying six actionable challenges for advocates of neural-based systems. The hope is that this article provides clear goals for those developing LRMTS to achieve in a similar way – a challenge list. This section will survey a few papers that take a similar approach.

Deployment. A number of papers have discussed **minimum viable product (MVP)** in the context of LRMT. (Joshi et al., 2019) discuss a number of challenges associated with creating systems for low-resource language communities. (Farajian et al., 2017) focus on the challenges of deployment related to multiple domains, a challenge covered later in Section 6.2. Their work discusses accuracy and other preconditions for deployment. (Garcia et al., 2023) comment on the effects of few-shot learning in LRMT when translating Icelandic. Other work (González Rubio, 2014) assesses the quality of human effort as a metric for MT system deployment. Their work addresses a few of our concerns but this article combines several sub-challenges not covered by theirs into one (the deployment viability challenge). Other task-specific MT work (Lewis et al., 2011) provides a "Crisis Cookbook" of terminology for a deployed LRMTS in crisis situations but does not address issues from generic LRMTS. Lastly, one of the more important investigations (Diddee et al., 2022) sheds light on bloated models that use distillation as a form of compressing models in low-resource system deployment. Their work encroaches on the same path as this because it takes into account the deployment of systems that use LLMs for low-resource settings, by far the most popular approach in current times.

Quality. Much has been written about LRMT quality. Initial work (Schiaffino and Zearo, 2005) introduce indices and software that were promising and included both the MT system and the human; while their work is notable, we focus on the following seminal work. Mentioned before as a *human value* challenge resource, work by (Castilho et al., 2018) extends previous work (Way, 2018; Moorkens et al., 2018) by introducing a translation quality assessment metric that we use in this work along with other measurements. Other automated methods such as the one from (Specia et al., 2013; Specia and Shah, 2018) focus more on creating predictors for quality rather than the challenge of measuring human versus machine.

Evaluation. When it comes to evaluation for LRMT, an article of this nature could report on many. However, there are some main resources that are used in determining the challenges consisting of the following work. The default standard measurements which cover string-based and embedding-based methods are already mentiond by (Haddow et al., 2022): BLEU (Papineni et al., 2002), ChrF (Popović, 2015), BERTscore (Zhang et al., 2020), COMET (Rei et al., 2020b), BLEURT, (Sellam et al., 2020) and METEOR (Denkowski and Lavie, 2011). Several major LRMT projects like GOURMET (Birch et al., 2019), Google Research (Siddhant et al., 2020), FLORES (Guzmán et al., 2019; Goyal et al., 2022) and more (Isabelle et al., 2017) currently use the standard metrics. One previous investigation (Östling and Tiedemann, 2017) used BLEU (Papineni et al., 2002) to determine that 70k sentences was sufficient to provide decent quality for a neural LRMTS. The assumption from LRMT developers is that including humans is expensive and time-consuming avoiding inclusion of more human-like measurements such as adequacy (Doherty, 2018), HTER (Snover et al.,

2006), and fluency (Reeder, 2004). This article describes uses of those metrics along with the following others to help better overcome the evaluation challenge. For the bias and culture challenge, we report two evaluation frameworks used for guidance: WinoMT (Stanovsky et al., 2019; Stafanovičs et al., 2020) and MT-GenEval (Currey et al., 2022). For evaluating human parity value with LRMT, seminal work (Castilho et al., 2018; Way, 2018) provides insight into translations as a whole used a translation quality assessment. One quality metric used for evaluation by projects like the one from (Bayón and Sánchez-Gijón, 2019) called the Multidimensional Quality Metric (MQM) (Lommel et al., 2014) identifies errors from the wide range of possibilities mentioned. However, it does not seem to take into account *bias and culture*, something that we address in this article.

Bias and Culture. A challenge only slightly investigated in the past, accountability of bias and culture has been identified as lacking in several sub-fields of NLP including MT and more specifically LRMT. Work done in 2020 (Hovy et al., 2020) has already shown that three commercial machine translation systems (Bing, DeepL, Google) have some sort of demographic bias in the training data. The evidence is further corroborated by other investigators in the field. For commercial systems, (Levy et al., 2021) have attempted to solve co-reference resolution pronouns and other gender bias. Another article (Haroutunian, 2022) has shown that LRMTS that do not collaborate with the end users can make communities vulnerable which addresses one of the major challenges when creating or deploying LRMTS. More published work (Stafanovičs et al., 2020) has mitigated bias by annotating words with gender information while others (Wang et al., 2021) sought out to explicitly include bias language for back and forward translation. Those efforts (Hovy et al., 2020; Stafanovičs et al., 2020; Wang et al., 2021; Haroutunian, 2022) have shown to be somewhat successful; but, they do not provide explicit thresholds to abide by. We feel that this article will help move their work in the right direction by providing thresholds and awareness as shown by (Daems and Hackenbuchner, 2022) who delivered a website[2] for detecting bias. Our work attempts to achieve results similar to (Drugan and Babych, 2010)'s work which provide clear direction as a guide of what one should do when creating an LRMTS. To achieve this, we use background work from (Saunders and Byrne, 2020) who used the WinoMT test (Stanovsky et al., 2019) and the MT-GenEval (Currey et al., 2022) framework as pre-cursors for writing Section 5.2.

Given the challenges of LRMT, it may be necessary to consider workarounds such as human-in-the-loop and subdomains. Much has been written about both of these subjects. Our discussion of human-in-the-loop follows Castilho et al. (Castilho et al., 2018). The main difference between their work and this article it that the topic is based on comparisons for quality alone, this article presents quality as a challenge but also presents other challenges, one of those being the value of a human-in-the-loop. Castilho et al. (Castilho et al., 2018) insights several of the key aspects and metrics such as *adequacy* and *fluency* that show the importance of a human in MT. The humans included in their project show that BLEU (Papineni et al., 2002) scores alone are not enough to judge LRMT output. We highlight their work in this article as a valid LRMT case for including humans. Other effects of human value are found in health crisis situations like natural disasters and more in Lewis et al. (Lewis et al., 2011) which provides direction on what key terminology to use for greater impact in times of crisis. Other evaluations (Haroutunian, 2022) construct *value scenarios* to create LRMTS as language-specific tools not language-agnostic ones. Their evaluations align closely with ours and should be considered an additional read when working with the LRMT challenges. Other broader work similar to this article yet not focused solely on LRMT is the work from Bender et al. (Bender et al., 2021) that recommends involving stakeholders (humans) when deploying systems backed by LLMs. Their work is closely related to our work but broader; however, it should be considered as a key piece of inspiration for this article.

Domain Specificity. We will discuss subdomains in Section 6.2. Some papers (Li et al., 2019; Moslem et al., 2023) attempt to solve the known domain problem via real-time adaptation techniques while other papers (Britz et al., 2017) use multiple domains in the same MT system. The domain challenge is obviously one of this most important challenges; in this article, we do not attempt to solve it, merely we attempt to provide baseline advice as to what should be accomplished. To do so,

[2] https://artificiallycorrec.wixsite.com/biasbyus

we rely on previous work (Haddow et al., 2022; Kreutzer et al., 2022) that notes that scarce data along with domain-specific LRMTS are a challenge. Additionally, they note that zero-shot or few-shot low-resource language model can worsen the problem. A good example of how a deployed LRMTS does not work well with multiple domains is the *human value* where standard biblical data (Agić and Vulić, 2019) did not perform well on everyday magazine data. Even more LRMT work (Ortega et al., 2021; Soto et al., 2022) gives proof on the challenges of translating source sentences in two languages (a low-resource language and its high-resource neighbor's language) to a domain-specific target language like clinical text or everyday prose.

5 Quality

The quality expectations of a LRMTS should be similar to those of a professional translator. An unfortunate by-product of the increasing amount of digital resources available is that they dampen performance due to higher search spaces. We consider the following attributes of a high-quality translation for different domains as highly important: (1) verified by humans and (2) adjusted to their domain (3) free of bias and (4) evaluated for accuracy. There are several techniques to guarantee quality of which the main two methods are: involving humans and estimating quality. Quality estimation of machine translation *must* have used a human-in-the-loop regardless if it is for the ground truth translations or the approval of MT system suggestions. We highly recommend the use of a framework such as the Translation Quality Assessment framework (Castilho et al., 2018) which should include several of the metrics mentioned in Section 5.1.

What determines if LRMTS translations are of high quality? Generally speaking, humans determine whether or not a translation is of high quality. Of course, in a LRMTS, the quality expectation are generally lower since most LRMTS do not tend to be of high quality. One way of measuring is called the translation edit rate (TER) (Snover et al., 2006) and it is the amount of edits that a professional translator would take for improving it. As for an acceptable TER score, acceptable ranges from previous work (Tonja et al., 2023; Denkowski and Lavie, 2010; Snover et al., 2006) for LRMTS should be $\approx$ 50–70 and by no means should they be more than 90 (a near useless translation). Other metrics such as HTER (Human TER), METEOR (Denkowski and Lavie, 2011), and BLEU (Papineni et al., 2002) are considered correlationary with humans and discussed further in Section 5.1.

Are there methods for estimating quality in a LRMTS without a human? Although there are automated methods for estimating the quality of an LRMTS, the methods generally use some form of reference (ground-truth) data as is the case of QuEST (Specia et al., 2013), a framework that uses word and sentence-level features for estimating quality similar to a human. We discourage the use of quality estimation and other automated techniques during the initial phases of the creation of a LRMTS that is intended to be deployed in the wild. As mentioned in this article, a human should always be involved despite the higher time and expense required, this is even more important during the initial development stage.

Can a machine determine LRMT quality better than a human? Simply put, there is not substitute for a human in the LRMT creation loop. At this point in time, to our knowledge, there does not exist a LRMTS that has achieved nearly the same performance as high-resource language pairs like English–German. While some BERT-based (Devlin et al., 2019) MT systems that use transformers (Vaswani et al., 2017) have achieved near-human performance when measured by BLEU (Papineni et al., 2002), it is not clear that is the case for LRMTS or domain-specific situations as was shown in recent work (Au Yeung et al., 2023) in the clinical domain.

5.1 Evaluation: What is "Good Enough"?

Several methods have been discussed in this article for evaluating LRMTS. SOTA review (Freitag et al., 2022) has shown that conventional methods such as BLEU (Papineni et al., 2002), COMET (Rei et al., 2020a) and CHRF (Popović, 2015) are not the best methods for neural LRMTS. Evaluation metrics for LRMTS should be a combination of the metrics introduced here and account for fluency, adequacy, human value, bias, and more. A diverse set of expectations is taken into account using the Multidimensional Quality Metric (MQM) (Lommel et al., 2014). We propose a comprehensive list of acceptable or typical ranges for deployable LRMTS below omitting those that we have already covered. Keep in mind, that the list is by no means exhaustive; additionally, major corporations have

already deployed several LRMTs for low-resource languages like Quechua and Basque with scores for these metrics that are lower. The assumption is that the LRMTS has a reasonable amount of data (more than 10k parallel sentences).

Metric	Range
BLEU (Papineni et al., 2002)	≈ 15–35
ChrF (Popović, 2015)	≈ 40–70
BERTscore (Zhang et al., 2020)	≈ 60–80
COMET (Rei et al., 2020b)	≈ 15–60
BLEURT (Sellam et al., 2020)	≈ 25– 50
METEOR (Denkowski and Lavie, 2011)	≈ 20–50
Fluency (Reeder, 2004)	≈ 1.0–3.0

Table 1: Typical Quality LRMT Metrics

The metrics and accompanying scores in Table 1 are meant to serve as a guide for what a company could expect from a LRMTS given the current systems that have been deployed in the wild. Most LRMTS are not good enough to use in the eyes of the low-resource community (Mager et al., 2023) but deployment can be considered for some cases like crises or others (O'Brien and Cadwell, 2017) as long as the proper care is taken to set appropriate expectations (especially for non-critical situations).

5.2 Bias and culture

One source of bad outputs are biases. Much has been written about biases and other risks (Savoldi et al., 2021; Bender et al., 2021; Church et al., 2022; Garcia et al., 2023). There are additional concerns for LRMT (Haroutunian, 2022), though there are also benefits, as discussed in Bird's TED Talk[3] as well as his keynote at ACL-2022[4]. Bird encourages us to treasure languages and stories (like gold); we should embrace diversity, and avoid patronizing/disrespectful terms (e.g., endangered, indigenous, ethnic). Hopefully, the benefits outweigh the risks.

6 Plan B: Workarounds

Given the realities of LRMT, it may be necessary to consider various workarounds in order to achieve quality that is good enough to deploy a minimum viable product. The next two sections consider two workarounds of many possible: (1) human-in-the-loop and (2) subdomains.

[3] https://www.youtube.com/watch?v=vfMIWqflNgE

[4] https://www.2022.aclweb.org/keynote-speakers

6.1 Plan B: Human-in-the-Loop

The high value of human annotation has already been shown in previous work. While claims are made by recent literature (Goyal et al., 2022) that a human's involvement is timely and expensive, it cannot be absent. In order to determine acceptable values for human involvement, we rely on the past investigation in the area (Koehn, 2009; González Rubio, 2014; Way, 2018; Castilho et al., 2018; Kreutzer et al., 2022; Saldías et al., 2022) to answer the main questions below.

How many human evaluators should a LRMTS include? While it should be clear that some human evaluation of the translation output from a LRMTS is better than none, effective LRMTS generally use more than one native human evaluator. For example, (Kumar et al., 2021) were able to show that despite BLEU (Papineni et al., 2002) scores around 8, general fluency was achieved when reviewed by 2 native speakers. Crowdsourcing on the internet provides another advantage to gain more annotators; however, (Persaud and O'Brien, 2019) have shown that the quality may be inferior to having human annotation in the project. Therefore, it is our suggestion that the LRMTS be evaluated by at least one native speaker with the ideal number of annotators (near-native or native) being from 3 to 5 given that the evaluation set is not terribly time-consuming or large (see work from (Castilho et al., 2018) for more details) and that the inter-annotator agreement (IAA) have a KAPPA coefficient range from 50 to 90%. (Birch et al., 2016; Bojar et al., 2016)

What metrics should a LRMTS use to measure a human's value? As previously mentioned, a high IAA is recommended. However, other metrics like quality of annotation and time taken should be considered. Resulting annotations, often times using an integral Likert scale like 1–5, should coincide with the desired output requirements of metrics like adequacy, fluency, and more (see Section 5.1 for suggested metrics). Previous work from (Kreutzer et al., 2022) measures IAA and uses non-native speakers for quality annotations – this provided evidence that it is not necessary to include all native speakers but IAA should be high. Other work (Castilho et al., 2018; Doherty, 2018) mentions that translation quality assessments around 60 to 70% are acceptable. For a LRMTS, the human involvement can lead to high quality LRMTS as shown by (Saldías et al., 2022).

6.2 Plan B: Subdomain

One of the major challenges for LRMTS is creating a multi-domain system that works well across broad states of categories. As an addition, a LRMTS could include several languages much like the work from (Guzmán et al., 2019). The expectations from our standpoint of view are two-fold: (1) the LRMTS should contain the maximum amount of parallel sentences available from varied sources and (2) the LRMTS should notify the user (allbeit an investigator or low-resource community user) of the intended domain (unless it is intended for the generic domain).

How many domains should a LRMTS target? The simple solution is that an LRMTS should target infinite domains; but, there is little research that shows this is possible. Other work (Chu and Wang, 2018; Zeng et al., 2018; Liang et al., 2021; Li et al., 2019; Moslem et al., 2023) explores the possibility using domain-adaptation techniques. We suggest that the LRMTS have a rapid way of protyping domain-specific cases like the work from (Palmer et al., 1998). While their work is nearly 30 years past, there is an important takeaway: they used a six-month effort with two native speakers (French and Arabic) to extend a generic domain to two specific domains in turn making the quality of both domains much better. While we cannot quantify the amount of resources that a LRMTS has on hand, we can use previous research as a way of suggesting that a continuous human-in-the-loop feedback development can be rewarding. This was also shown in recent work for low-resource Irish in the Covid domain (Lankford et al., 2021) – they achieved improvements of 27 points in BLEU (Papineni et al., 2002) with 5,000 high-quality translations that included human evaluation.

Should the LRMTS mix training data? There is no simple answer to this question. However, a LRMTS developer could take into account the amount of resource available to determine what would be best. For example, in parallel corpora benchmarks like Flores (Goyal et al., 2022) with around 200,000 parallel sentences on multiple domains achieve $\approx$ 10 BLEU (Papineni et al., 2002). Unless created for a crisis situation, this system would probably not be deployable. However, for domain-specific purposes like law or medicine, if 200,000 parallel sentences were available, a SOTA technique (Reheman et al., 2023) can achieve reasonable BLEU (Papineni et al., 2002) scores making it viable. Therefore, it is our suggestion that LRMTS would be better off if they have domain-specific parallel data on the order of hundreds of thousands.

What can be done to overcome the lack of data in multiple domains? As previously stated, if data does not exist in a domain, one of the most viable options would be a domain-adaptation technique that includes native speakers and human evaluation for feedback. In Section 5, we discuss how quality should be measured. There is no doubt that this would be time-consuming but we disagree that "some change is better than no change" (Wagstaff, 2012). Since systems generally do not achieve the quality necessary to be deployed in non-crisis situations, when native speakers and others are not available to verify adaptation or augmentation techniques, we feel that it is best not to create or deploy the LRMTS.

7 Conclusions

We provided a practical guide of challenges for companies considering the deployment of a LRMTS. Much of the work in our field focuses on English and other high-resource language, but recently, there has been more interest in low-resource languages. A number of systems support an amazingly large set of languages. That said, it is a mistake to deploy a non-viable system. Adoption of LRMT can be limited by many factors and the question therein lies if the risks are worth the rewards. A company's minimum viable product requires sufficient demand with hundreds (if not thousands) of users in the low-resource community that are willing to use it. In addition to demand, we also discussed costs and quality. Quality includes standard metrics in Table 1 such as BLEU (Papineni et al., 2002), as well as other considerations such as bias and respect for cultural diversity. These other considerations may be more difficult to quantify, but that should not diminish their value. In his TED Talk, Bird (Footnote 3) encourages us to treasure languages and stories (like gold); we should embrace diversity, and avoid patronizing/disrespectful terms (e.g., endangered, indigenous, ethnic). Quality tends to increase with the size of the training set. For low resource languages, it may not be feasible to improve quality by increasing the size of the training set. Two workarounds were discussed to address these realities: (a) human-in-the-loop and (b) subdomains.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 3204–3210.

Joshua Au Yeung, Zeljko Krajevic, Alfred Balston, and James T ..., Teo. 2023. Ai chatbots not yet ready for clinical use. *medRxiv*, pages 2023–03.

Kalika Bali, Monojit Choudhury, and Vivek Seshadri. 2019. Ellora: Enabling low resource languages with technology. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 160–163.

María Do Campo Bayón and Pilar Sánchez-Gijón. 2019. Evaluating machine translation in a low-resource language combination: Spanish-galician. In *Proc. of the MT Summit XVII: Translator, Project and User Tracks*, pages 30–35.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proc.of the 2021 ACM Conf. on Fairness, Accountability, and Transparency*, pages 610–623.

Alexandra Birch, Omri Abend, Ondrej Bojar, and Barry Haddow. 2016. Hume: Human ucca-based evaluation of machine translation. *arXiv preprint arXiv:1607.00030*.

Alexandra Birch, Barry Haddow, Ivan Titov, Juan Antonio ..., Pérez-Ortiz, et al. 2019. Global under-resourced media translation (gourmet). In *MTSummit (2)*, page 122.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of wmt evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proc. of the 2nd Conference on Machine MT*, pages 118–126. ACL.

Patrick Cadwell. 2021. Translation and interpreting in disaster situations. *The Routledge Handbook of Translation and Health*, pages 253–268.

Luis Camacho and Rodolfo Zevallos. 2020. Language technology into high schools for revitalization of endangered languages. In *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4. IEEE.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In *Translation quality assessment*, pages 9–38. Springer.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*.

Kenneth Church, Annika Schoene, John E. Ortega, Raman Chandrasekar, and Valia Kordoni. 2022. Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable. *Natural Language Engineering*, page 1–26.

Marta R Costa-jussà, James Cross, Onur Çelebi, Jean ..., Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, and Georgiana ..., Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proc. of the 2022 Conf. on EMNLP*, pages 4287–4299.

Joke Daems and Janiça Hackenbuchner. 2022. DeBiasByUs: Raising awareness and creating a database of MT bias. In *Proc. of the 23rd Annual Conference of the EAMT*, pages 289–290.

Michael Denkowski and Alon Lavie. 2010. Extending the meteor machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conference of NAACL*, pages 250–253.

Michael J. Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *WMT@EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.

Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, and Kalika Bali. 2022. Too brittle to touch: Comparing the stability of quantization and distillation towards developing low-resource MT models. In *Proceedings of the 7th Conference on MT (WMT)*, pages 870–885. ACL.

J. M. Doherty. 2018. Translation quality assessment. In *Machine Translation: Technologies and Applications*.

Jo Drugan and Bogdan Babych. 2010. Shared resources, shared values? ethical implications of sharing translation resources. In *Proc. of the 2nd Joint EM+/CNGL Workshop*, pages 3–10.

Abteen Ebrahimi, Arya D McCarthy, Arturo Oncevay, and Katharina ..., Kann. 2023. Meeting the needs of low-resource languages: The value of automatic alignments via pretrained models. *arXiv preprint arXiv:2302.07912*.

M Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. Neural vs. phrase-based machine translation in multi-domain scenario. In *Proc. of the 15th Conference of the EACL: Volume 2, Short Papers*, pages 280–284. The ACL.

Markus Freitag, Ricardo Rei, Nitika Mathur, and André FT ..., Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the 7th Conference on MT (WMT)*, pages 46–68.

Xavier Garcia, Yamini Bansal, Colin Cherry, and Orhan ..., Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *arXiv preprint arXiv:2302.01398*.

Jesús González Rubio. 2014. *On the effective deployment of current machine translation technology*. Ph.D. thesis, Universitat Politècnica de València.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, and Angela ..., Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *ACL*, pages 522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, and Aurelio ..., Marc. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proc. of the 2019 Conf. on EMNLP-IJCNLP*, pages 6098–6111.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, pages 673–732.

Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proc. of the 60th Annual Meeting of the ACL: Student Research Workshop*, pages 44–54.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proc. of the 58th Annual Meeting of the ACL*, pages 1686–1690.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proc. of the 2017 Conference on EMNLP*, pages 2486–2496.

Haukur Pall Jonsson, Haukur Barri Simonarson, Vesteinn Snbjarnarson, Steinor Steingrimsson, and Hrafn Loftsson. 2020. Experimenting with different machine translation models in medium-resource settings. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*, pages 95–103.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Markus ..., Freitag, et al. 2022. Proceedings of the 7th conf. on mt (wmt). In *Proceedings of the 7th Conf. on MT (WMT)*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, and Mofetoluwa ..., Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the ACL*, pages 50–72.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797*.

Seamus Lankford, Haithem Afli, and Andy Way. 2021. Machine translation in the covid domain: an english-irish case study for loresmt 2021. In *Proceedings of the 4th Workshop on LORESMT*, pages 144–150.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the ACL: EMNLP 2021*, pages 2470–2480.

William Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis mt: Developing a cookbook for mt in crisis situations. In *Proc. of the 6th Workshop on SMT*, pages 501–511.

Rumeng Li, Xun Wang, and Hong Yu. 2019. Metamt, a metalearning method leveraging multiple domain data for low resource machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8245–8252.

Jianze Liang, Chengqi Zhao, Mingxuan Wang, Xipeng Qiu, and Lei Li. 2021. Finding sparse structures for domain specific neural machine translation. In *Proceedings of the AAAI Conference on AI*, pages 13333–13342.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, pages 0455–463.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. *arXiv preprint arXiv:2305.19474*.

Marion Weller-di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proc. of the 7th Conference on MT (WMT)*, pages 801–805.

Gary Marcus and Ernest Davis. 2020. Gpt-3, bloviator: Openai's language generator has no idea what it's talking about. *Technology Review*.

Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. 2018. Translation quality assessment. *Machine translation: Technologies and applications ser. Cham: Springer International Publishing*, 1:299.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors. 2022. *Proc. of the 60th Annual Meeting of the ACL (Vol.1: Long Papers)*.

Sharon O'Brien and Patrick Cadwell. 2017. Translation facilitates comprehension of health-related crisis information: Kenya as an example. *Journal of Specialised Translation*, 1(28):23–51.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, pages 325–346.

John E Ortega, Iria de Dios-Flores, Pablo Gamallo, and José Ramom Pichel. 2022. A neural machine translation system for galician from transliterated portuguese text. In *Proceedings of the Annual Conference of the SEPLN. CEUR Workshop Proceedings*, pages 92–95.

John E Ortega, Richard Alexander Castro Mamani, and Jaime Rafael Montoya Samame. 2021. Love thy neighbor: Combining two neighboring low-resource languages for translation. *Proceedings of MT Summit XVIII*.

Robert Östling and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *arXiv preprint arXiv:1708.05729*.

Martha Palmer, Owen Rambow, and Alexis Nasr. 1998. Rapid prototyping of domain-specific machine translation systems. In *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas AMTA'98 Langhorne, PA, USA, October 28–31, 1998 Proceedings 3*, pages 95–102. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th annual meeting of the ACL*, pages 311–318.

Ajax Persaud and Steven O'Brien. 2019. Quality and acceptance of crowdsourced translation of web content. In *Social Entrepreneurship: Concepts, Methodologies, Tools, and Applications*, pages 1177–1194. IGI Global.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proc. of the 10th workshop on SMT*, pages 392–395.

Florence Reeder. 2004. Investigation of intelligibility judgments. In *Conference of the AMTA*.

Abudurexiti Reheman, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. *arXiv preprint arXiv:2301.05380*.

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

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020b. Comet: A neural framework for mt evaluation. In *Conference on EMNLP*.

Belén Saldías, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. *arXiv preprint arXiv:2204.05307*.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proc. of the 58th Annual Meeting of the ACL*, pages 7724–7736. ACL.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the ACL*, pages 845–874.

Riccardo Schiaffino and Franco Zearo. 2005. Translation quality measurement in practice. In *Proc. of the 46th Annual Conference of the AMTA*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Annual Meeting of the ACL*.

Aditya Siddhant, Melvin Johnson, Henry Tsai, and Karthik ..., Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *Proc.of the AAAI conference on AI*, pages 8854–8861.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of the 7th Conference of the AMTA: Technical Papers*, pages 223–231.

Xabier Soto, Olatz Perez-de Viñaspre, Maite Oronoz, and Gorka Labaka. 2022. Development of a machine translation system for promoting the use of a low resource language in the clinical domain: The case of basque. In *NLP in Healthcare*, pages 139–158. CRC Press.

Lucia Specia and Kashif Shah. 2018. Machine translation quality estimation: Applications and future perspectives. *Translation quality assessment: from principles to practice*, pages 201–235.

Lucia Specia, Kashif Shah, José GC De Souza, and Trevor Cohn. 2013. Quest-a translation quality estimation framework. In *Proc. of the 51st Annual Meeting of the ACL: System Demonstrations*, pages 79–84.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the 5th Conf. on MT*, pages 629–638.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proc. of the 57th Annual Meeting of the ACL*, pages 1679–1684.

Maali Tars, Taido Purason, and Andre Tättar. 2022. Teaching unseen low-resource languages to large translation models. In *Proceedings of the 7th Conference on MT (WMT)*, pages 375–380. ACL.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2023. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, page 1201.

Ashish Vaswani, Noam Shazeer, Niki Parmar, and Illia ..., Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*.

Kiri Wagstaff. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656*.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, and Yang ..., Liu. 2021. On the language coverage bias for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4778–4790, Online. Association for Computational Linguistics.

Andy Way. 2018. Quality expectations of machine translation. *Translation quality assessment: From principles to practice*, pages 159–178.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proc. of the 2018 Conference on EMNLKP*, pages 447–457.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.