Towards Multilingual Automatic Open-Domain Dialogue Evaluation

John Mendonça[1,2,*], **Alon Lavie**[3,4] and **Isabel Trancoso**[1,2]
[1] INESC-ID, Lisbon
[2] Instituto Superior Técnico, University of Lisbon
[3] Carnegie Mellon University, Pittsburgh
[4] Phrase, Pittsburgh
{john.mendonca, isabel.trancoso}@inesc-id.pt
alavie@cs.cmu.edu

Abstract

The main limiting factor in the development of robust multilingual open-domain dialogue evaluation metrics is the lack of multilingual data and the limited availability of open-sourced multilingual dialogue systems. In this work, we propose a workaround for this lack of data by leveraging a strong multilingual pretrained encoder-based Language Model and augmenting existing English dialogue data using Machine Translation. We empirically show that the naive approach of finetuning a pretrained multilingual encoder model with translated data is insufficient to outperform the strong baseline of finetuning a multilingual model with only source data. Instead, the best approach consists in the careful curation of translated data using MT Quality Estimation metrics, excluding low quality translations that hinder its performance.

1 Introduction

Open-domain dialogue systems have gained substantial attention in the NLP (Natural Language Processing) and ML (Machine Learning) fields, thanks to their increasingly human-like behaviour (Thoppilan et al., 2022; Shuster et al., 2022). Their impressive generation capabilities can be attributed to new milestones in model development and scaling (Adiwardana et al., 2020), and the amount of data used during training. Despite this research and development effort, advertised generation capabilities were only attainable in a select few languages (typically English or Chinese) due to low resources in dialogue for other languages (Zhang et al., 2022b). More recently, however, the advent of LLMs (Large Language Models) finetuned with Reinforcement Learning from Human Feedback such as ChatGPT (Ouyang et al., 2022) has opened the path for high-quality and easily accessible multilingual dialogue generation.

Similarly, automated open-domain dialogue evaluation has also been largely limited to evaluating a select few languages. Word-overlap based metrics from NLG (Natural Language Generation) such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are agnostic to language, only requiring a reference response. However, these metrics are known to correlate poorly with human judgments due to the multifaceted nature of dialogue (Liu et al., 2016). Reference-free metrics such as USR (Mehri and Eskenazi, 2020) and USL-H (Phy et al., 2020), however, require dialogue data for training. Considering most open-source dialogue data is in English, these models are expected to underperform significantly in other languages. Additionally, most open sourced dialogue systems are also limited to English, further disincentivising multilingual research.

One solution to the issues previously mentioned is to leverage MT (Machine Translation). With MT services becoming more affordable and consistent, some authors resort to translation when developing their multilingual dialogue systems (Schuster et al.,

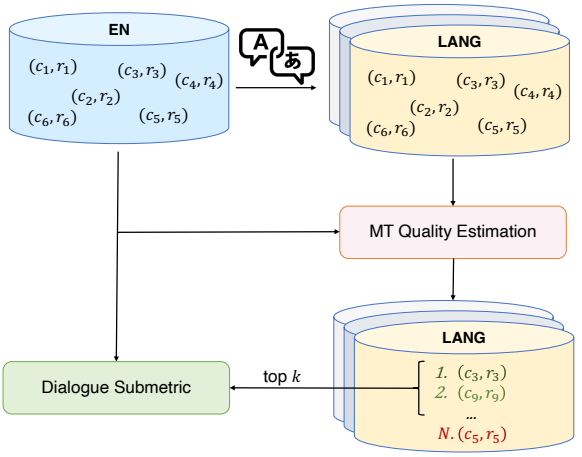


Figure 1: Proposed architecture. The original dialogue dataset is transformed into context-response pairs (c_n, r_n) and translated using MT. The final dialogue submetric is trained using a combination of the original English data and the top k sentences or (c_n, r_n) from each language, depending on the submetric.

[*] Work conducted as a visiting scholar at CMU.

2019; Anastasiou et al., 2022). This can either be included as a module in the system's pipeline – allowing the use of proven English generation models for other languages; or as a cross-lingual transfer method – by translating training data.

In this paper, we extend the approach of training using data generated by MT for the development of multilingual models for evaluation of open-domain dialogue responses. We experiment with and evaluate several different possible workarounds for this problem. Namely, we leverage the availability of strong pretrained multilingual encoders as a foundation for training multilingual dialogue evaluation models. As a first step, we translate existing publicly-available English dialogue data into the target languages. We then explore multiple alternative ways to leverage this translated data in order to finetune and train monolingual and multilingual dialogue evaluation models for two specific dialogue submetrics. To address the impact of low quality translations, we propose using an MT Quality Estimation (QE) model to rank the translations and investigate the impact of finetuning models with varying amounts of quality-ranked data. Figure 1 illustrates the proposed approach.

The performance of these alternative models is evaluated on a curated test set of dialogues which were human-annotated with dialogue quality scores for two subqualities. The original English test set was translated using MT and then post-edited by editors into six different target languages (PT-Portuguese, DE-German, FR-French, ZH-Chinese, ES-Spanish and JA-Japanese). The quality scores from the human annotations of the original English dialogues were then carried over to the target-language dialogues. Our finetuned multilingual dialogue evaluation models exhibit strong correlations with human judgements, comparable to LLMs, indicating it is possible to leverage multilingual dialogue evaluation metrics without the constraints LLMs currently possess (costs, latency, etc.). We hope this will encourage other researchers to update existing metrics using our proposed multilingual finetuning approach.

In summary, the primary contributions of this work are as follow:

- We evaluate cross-lingual transfer and translation augmented training approaches using MT for the task of training multilingual dialogue evaluation models, showing that, on average, the best performance is achieved by finetuning with subsets consisting of only the best translations. We found that, depending on the subquality and target language, the optimal amount of translated data can be as low as 5% and as high as 75%.
- We translate and release DailyDialog and a corresponding test set of human quality annotations in 6 languages to facilitate future benchmarking of multilingual dialogue evaluation metrics[1].

2 Background

2.1 Open-Domain Dialogue Evaluation Metrics

The recent trend in open-domain dialogue evaluation is to train dialogue submetrics using well-defined self-supervised tasks which correlate well with their corresponding subqualities. The most used self-supervised task is Next Sentence Prediction (NSP), as it is known to correlate well with subqualities that evaluate *"Context Awareness"*. Examples of this include: *Uses Context* (Mehri and Eskenazi, 2020), *Sensibleness* (Phy et al., 2020; Mendonca et al., 2022) and *Relevance* (Zhao et al., 2020; Zhang et al., 2022a). Other subqualities include: *Fluency, Grammatically Correct* or *Understandability*, which use word-level noising techniques to generate negative samples (Phy et al., 2020; Mendonca et al., 2022; Zhang et al., 2022a); and *Specificity*, which uses an MLM (Masked Language Modelling) score (Mehri and Eskenazi, 2020; Phy et al., 2020; Zhang et al., 2022a). For overall quality, these submetrics are typically combined using different methods (e.g. empirical observation, trained Linear Regression or multilayer perceptrons).

To the best of our knowledge, there has not been any published research on cross-lingual transfer and/or development of trained multilingual metrics for open-domain dialogue evaluation.

2.2 Multilingual Text Classification

Despite the lack of research on multilingual dialogue evaluation, extending text classification to other languages is a well established subfield of research in NLP. The main constraint for multilingual performance parity is the lack of task-specific resources in the vast majority of written languages. Given the creation of these resources is

[1] github.com/johndmendonca/DialEvalML

both time consuming and expensive, most research effort has been geared towards general-purpose cross-lingual representations that are learned in an unsupervised way, therefore leveraging the unstructured data available in the wild. Large multilingual Transformer-based models (e.g mBERT, XLM-RoBERTa, and mT5) have been successfully used in a variety of classification tasks (Conneau et al., 2020; Pires et al., 2019; Xue et al., 2021). The standard approach for cross-lingual transfer is to finetune on existing domain data in a source language and perform inference in a target language. However, this approach typically lags behind models specifically trained with in-domain (both task and language) data.

As a solution to this problem, Pfeiffer et al. (2020) propose learning language-specific adapter modules via MLM on unlabelled target-language data followed by task-specific adapter modules by optimising a target task on labelled data in the source language. Task and language adapters are stacked, allowing cross-lingual transfer to the target language by substituting the target-language adapter at inference.

Bornea et al. (2021) propose an augmentation strategy where a corpus of multilingual silver-labelled QA pairs is generated by combining the original English training data with MT-generated data. A language adversarial training and arbitration framework bring the embeddings closer to each other, making the model language invariant.

To the best of our knowledge, there has not been any research on the utilization of MT Quality Estimation (QE) scoring as a means for identifying and demoting or excluding poorly translated data in such cross-language training scenarios.

3 Problem Formulation

The goal of reference-free turn-level dialogue evaluation is, given a dialogue history (frequently denoted as context) c of varying amount of turns, and a response r, to learn a scoring function that assigns a score $f(c, r) \rightarrow s$. This scoring function is compared against human judgements, which annotate the same context-response pairs. These responses are evaluated using a scaling method, for instance, a binary $(0, 1)$ judgement or a $[1, 5]$ scale, where the lowest value means lowest quality and highest value maximum quality. The notion of quality varies wildly depending on the annotation. In this work, we evaluate dialogue in two dimensions:

- **Understandability** An understandable response is one that can be understood without context. Such responses may contain minor typos that do not hinder the comprehension of the response.

- **Sensibleness** A sensible response is one that takes into account its preceding context.

Most automatic evaluation metrics reformulate the problem as regression. Performance is then evaluated using Pearson and Spearman correlations with human annotations.

3.1 Automatic Dialogue Evaluation Metrics

The majority of competitive metrics for dialogue evaluation include models trained in a self-supervised way for Valid Sentence Prediction (VSP) and Next Sentence Prediction (NSP) (Yeh et al., 2021; Zhang et al., 2021). As such, the focus of this work was to evaluate multilingual dynamics for these models, which can then be employed on existing metrics.

VSP: Valid Sentence Prediction In this paper, we followed the approach used by Phy et al. (2020) and initially proposed by Sinha et al. (2020). A regression model was trained to differentiate between positive samples and synthetic negative samples. **Positive** samples are perturbed by randomly applying one of the following: (1) no perturbation, (2) punctuation removal, (3) stop-word removal. **Negative** samples are generated by randomly applying one of the following rules: (1) word reorder (shuffling the ordering of the words); (2) word-drop; and (3) word-repeat (randomly repeating words).

NSP: Next Sentence Prediction The task of predicting sensibleness can be considered a binary (NSP) task, distinguishing a positive example from a semantically negative one, given a context. A discriminative regression model was trained using the following sampling strategy: **positive** responses are drawn directly from the dialog; **negative** responses are randomly selected and a token coverage test discards semantically similar sentences. All responses are processed using the positive-sample heuristic used by VSP.

4 Cross-lingual Transfer Learning

The goal of the experiments described in this section was to evaluate different basic approaches of

cross-lingual transfer for the task of automatic dialogue evaluation. For encoder model training, we leveraged Machine Translation (MT) by fully translating an English source dialogue dataset and then finetuning monolingual and multilingual models using these translations.

4.1 Experimental Setup

4.1.1 Dataset

All experiments in this paper were based on the **DailyDialog (Li et al., 2017)** dataset, a high-quality human-human open-domain dialogue dataset focused on day-to-day conversations. After processing, we obtained train/dev splits of 58,515/25,078 and 89,707/38,449 per language for the VSP and NSP models, respectively. For training and evaluation, the post-processed dataset was translated into the target languages using MBART50 (Liu et al., 2020). We opted for using MBART50 as it is a relatively lightweight open sourced model with a large language coverage.

For the test set, we leveraged the annotations from Phy et al. (2020). These human annotations evaluate five responses from two retrieval methods, two generative methods, and one human-generated response for 50 contexts. These responses were annotated in terms of *Understandability* and *Sensibleness* [2]. We translated this set using Unbabel's[3] translation service. A total of 300 sentences were translated, corresponding to the 50 shared contexts and 250 responses. The translations were then split into smaller tasks and were corrected by editors from a commercial provider. Editors were specifically asked to retain any source disfluencies or hallucinations stemming from low quality response generation (e.g. *"I'm afraid you can't. I'm afraid you can't."*; *"Au contraire, you need to be a bahn."*). This ensured the original human quality annotations remained valid for the translation. A secondary senior editor reviewed the edited content as a whole.

4.1.2 Finetuned Encoders

We used XLM-RoBERTa (Conneau et al., 2020) as the encoder model for the experiments. This model is the multilingual version of RoBERTa, pretrained on CommonCrawl data containing 100 languages. For both the VSP and NSP models, we added a regression head on top of the encoder model.

EN – Zero-shot inference As a baseline for our results, we conducted zero-shot inference on the target languages using a model finetuned only on the original English data.

LANG – Target-Language Finetuning We finetuned the encoder with target-language translated dialogue data only. The downside of this approach is that a unique model needs to be trained for each target language. However, this method can be scaled to every language, including new ones, and is optimised to perform best in that language.

ML – Multilingual Finetuning Instead of finetuning a new model for each target language, one can finetune a single multilingual model by combining all of the translated data. In this case, the resulting single trained model is then used to evaluate responses in all languages. However, its performance may suffer in languages it has not seen during finetuning, even if they are supported by the encoder model. Furthermore, unlike target-language finetuned, the multilingual model is optimised jointly for all included languages.

MAD-X In this approach, we trained a VSP and NSP task adapter using the original English data by stacking the task adapter with a pretrained English language adapter (kept frozen during training). For zero-shot inference, the English language adapter was replaced by the target-language counterpart, while keeping the trained task adapter in place.

4.1.3 Large Language Model

As an additional strong baseline, we leveraged `gpt-3.5-turbo` (colloquially known as ChatGPT) as an evaluator of Understandability and Sensibleness. The context (exclusively for Sensibleness) and response was provided as input, together with the prompt *"{Given the context,} evaluate from 1-5 the response in terms of {dimension}. Provide the score and nothing else."*. This prompt, paired with a temperature setting of 0.0 attempted to minimises the variability of the output. Nevertheless, we report a standard deviation of (.003, .003) and (.001, .001) for Understandability and Sensibleness correlations, respectively, across 3 runs.

4.2 Results

The correlation results for all subqualities and the overall quality are presented in Table 1.

[2] Annotations for *Specificity* and *Overall Quality* were also conducted, but were excluded since they do not map to the learned metrics under study.

[3] unbabel.com

	EN		PT		DE		FR		ZH		ES		JA		AVG	
	Pr.	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**
Understandability																
EN	.376	.187	**.366**	.167	.328	.172	.351	.120	**.318**	**.202**	.342	.204	**.204**	.176	.327	.194
LANG	-	-	.176	.164	.214	.138	*.052*	.034	.274	.156	.219	.144	.185	.132	.214	.146
ML	.336	.117	.176	.167	.262	.150	*.012*	.015	.225	.138	.117	.158	*.091*	*.092*	.174	.126
MAD-X	.363	.166	.189	.103	.237	.122	.168	*.078*	.305	.168	.217	.119	.119	.129	.228	.126
ChatGPT	**.397**	**.334**	.365	**.230**	**.332**	**.263**	**.369**	**.273**	.276	.182	**.394**	**.263**	**.228**	**.223**	**.337**	**.263**
Sensibleness																
EN	.658	.676	.636	.651	.657	.655	.646	.656	.640	.656	.646	.657	.590	.599	.639	.649
LANG	-	-	**.649**	.661	.669	**.699**	.635	.655	.634	**.671**	.629	.669	.617	**.640**	.642	**.664**
ML	.651	.691	.606	**.675**	.634	.680	.605	**.669**	.642	.667	.596	.676	.599	.637	.619	**.664**
MAD-X	.660	.681	.614	.604	.664	.652	.624	.624	.608	.647	**.688**	.661	.558	.595	.631	.638
ChatGPT	**.746**	**.724**	.636	.626	**.683**	.675	**.695**	.666	**.655**	.645	.680	**.677**	**.625**	.610	**.674**	.662

Table 1: Average correlation results across 3 runs with different seeds. **Pr.** denotes Pearson and **Sp.** denotes Spearman. **Bold** denotes best performance, *Italic* $p < 0.05$.

Understandability The results show that, on average, the best performing encoder approach is the zero-shot inference using the English model (**EN**). Both the target-language finetuning (**LANG**) and multilingual finetuning approaches (**ML**) have much lower performances, indicating that translation augmentation is detrimental for this task. We also note that the **MAD-X** approach, although performing slightly better than ML and LANG, still lags behind EN considerably. In any case, ChatGPT largely outperforms other models on both metrics.

Sensibleness The best performing encoder approach for this subquality is LANG. Intuitively this makes sense, given that during finetuning the model is exposed to target-language data for the language it is being evaluated on. Furthermore, the performance difference between the different approaches is relatively much smaller, which indicates the Sensibleness subquality is less sensitive to MT quality. When comparing these results with ChatGPT, we observe a much smaller performance gap, with the best encoder models slightly outperforming on Spearman.

5 MT Quality-aware finetuning

The effects of noise introduced to the training data is a subject of intense research in the literature (Zhang et al., 2017; Hu et al., 2020; Swayamdipta et al., 2020). It is expected that, for this task, noise is introduced by low quality translations, reducing the performance of trained models. This issue was identified in Section 4, where for the VSP model in particular, the models trained using translations performed much worse than the baseline approach. Our hypothesis is that some translations heavily disrupt morphosyntactic cues used to infer response fluency, as shown in Table 2. We acknowledge that these low quality translations may also reduce

EN: Yes, I'd like to see the receipt.
Oh ! I see you bought the watch last week.
PT: Sim, gostava de ver o receio.
Oh! Vejo-te a fazer o relógio na semana passada.
QE score: -0.670

EN: Just look around ? Ah, that's boring.
ES: ¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡
QE score: -1.481

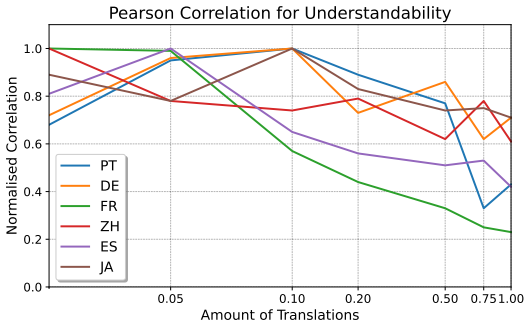
EN: Eight tens, six ones and large silver for others.
ZH: 八个十个,六个十个,其他十个十个十个十个十...
QE Score: -1.312

Table 2: Examples of low quality translations with corresponding QE score. Red denotes MT error, with underline in the source sentence indicating the closest alignment of the error. Blue denotes keywords that refer to prior context.

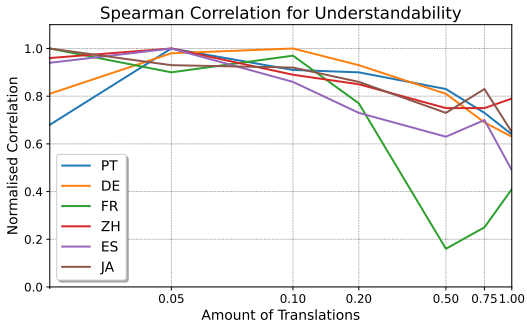
the quality of the response by disrupting keywords that point to the context (which is important for Sensibleness), or even more subtle quality cues (e.g. loss of empathy, inconsistency with named entities). However, the NSP model is trained to discriminate between the original response and randomly selected response from the corpus. As such, the model's prediction will remain invariant to most translation errors.

These observations, paired with the fact encoder models only slightly underperform ChatGPT (a much larger and expensive model), motivate the work described in this section. We hypothesise that, by ameliorating the MT noise via identifying and filtering low quality translations, the encoder model performance can outperform LLMs such as ChatGPT, at a fraction of the cost.

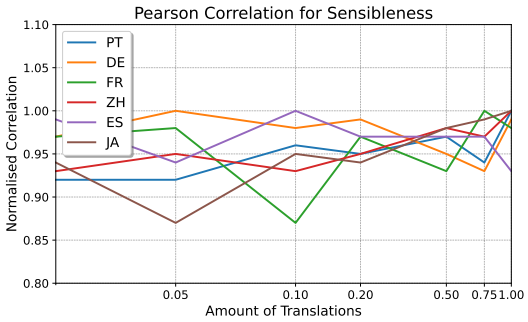
Since there are no available references, an MT QE (Specia et al., 2018) automatic metric is used for this purpose. Formally, an MT QE model is a scoring function that assigns a score given a source sentence and hypothesis translation. The unboundedness and uncalibrated nature of this score across languages results in the need for a cumbersome



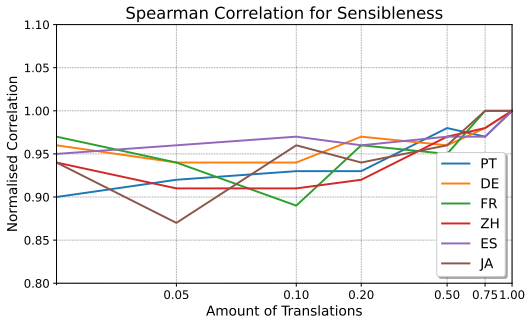
(a) Pearson Correlation, Understandability.



(b) Spearman Correlation, Understandability.



(c) Pearson Correlation, Sensibleness.



(d) Spearman Correlation, Sensibleness.

Figure 2: Normalised Pearson and Spearman correlation for the Understandability and Sensibleness submetric with varying amount of translated training data. Numeric results available in Appendix B.

analysis for each individual language in order to determine a threshold for filtering. Instead, we propose to use QE scores for response ranking, for each target language. This ensures a standardised method for filtering, improving the scalability of this method to new languages.

5.1 Experimental setup

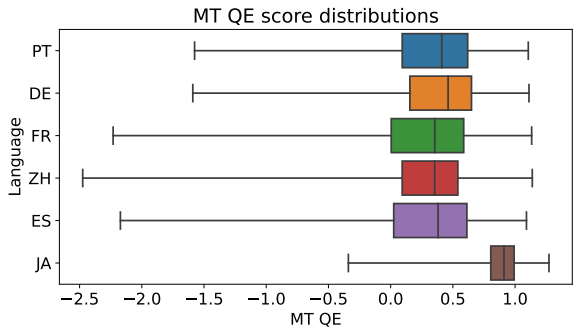


Figure 3: MT QE unnormalised score boxplot per language.

In order to confirm our hypothesis, we retrained all models using different amounts of translated data (100, 75, 50, 20, 10 and 5%). The ranking of the translations was conducted by scoring them using the WMT20 COMET-QE-DA model (Rei et al., 2020). For the VSP model, we ranked the individual sentences, and then applied negative sampling. For the NSP model, we ranked the positive and negative samples separately and then merged them together. Figure 3 presents the unnormalised score boxplot per language for all sentences (context and responses) for DailyDialog.

One of the things we noticed when finetuning the monolingual models was that the VSP models had large variations in performance. This can be attributed to (1) the low amount of training data, especially when using very few examples (5%, 10%), and (2) low quality translations, which is the research question this experiment attempts to answer. Since the true impact of low quality translations is obfuscated by other factors, we decided to finetune the LANG models starting from the EN checkpoint instead of the pretrained XLM-RoBERTa, and include the zero-shot results as 0%.

5.2 Results

LANG For the monolingual models, we plot normalised correlation results with the amount of MT data used during finetuning in Figure 2. The ***Understandability*** correlation results show that the optimal amount of translated data is language dependent, but with a clear indication that the inclu-

	EN		PT		DE		FR		ZH		ES		JA		AVG	
	Pr.	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**
Understandability																
0 (EN)	.376	.187	.366	.167	.328	.172	.351	.120	.318	.202	.342	.204	.204	.176	.327	.194
5	.403	.182	.490	.219	.344	.172	**.385**	*.091*	.320	**.235**	**.429**	.236	**.230**	.179	**.372**	.211
10	.377	.180	**.514**	.227	**.381**	.193	.294	*.091*	**.338**	.214	.385	.212	.216	.175	.358	.206
20	.384	.177	.478	.236	.333	.203	.153	*.087*	.318	.219	.315	.214	.174	.168	.308	.202
50	**.413**	.201	.481	**.242**	**.381**	.213	*.103*	*.053*	.310	.200	.315	.221	.219	.149	.317	.200
75	.311	.145	.247	.211	.320	.195	*.047*	*.048*	.163	.149	.111	.198	*.108*	.127	.187	.158
100	.336	.117	.176	.167	.262	.150	*.012*	*.015*	.225	.138	.117	.158	*.091*	*.092*	.174	.126
ChatGPT	.397	**.334**	.365	.230	.332	**.263**	.369	**.273**	.276	.182	.394	**.263**	.228	**.223**	.337	**.263**
Sensibleness																
0 (EN)	.658	.676	.636	.651	.657	.655	.646	.656	.640	.656	**.646**	.657	.590	.599	.639	.649
5	.637	.674	.629	.632	.627	.648	.637	.656	.629	.646	.626	.647	.567	.596	.621	.640
10	.642	.675	.639	.664	.661	.669	.636	.661	.637	.656	.635	.668	.575	.604	.632	.654
20	.650	.689	.627	.670	.649	.681	.627	.666	.621	.661	.637	.673	.568	.614	.626	.660
50	.667	.691	**.642**	.687	.650	.672	.621	.662	.652	.664	.629	.673	.600	**.642**	.637	.666
75	.677	.712	.629	**.694**	.679	**.702**	.633	**.679**	**.661**	**.673**	.643	**.695**	.593	.635	.645	**.679**
100	.651	.691	.606	.675	.634	.680	.605	.669	.642	.667	.596	.676	.599	.637	.619	.664
ChatGPT	**.746**	**.724**	.636	.626	**.683**	.675	**.695**	.666	.655	.645	.680	.677	**.625**	.610	**.674**	.662

Table 3: Average correlation results across 3 runs with different seeds for multilingual models when varying the amount of translated data.

sion of more translations decreases performance significantly. Instead, a lower amount of translations (5-10%) yields optimal performance. This shows that this small finetuning step is essentially adapting a model that was already finetuned for the downstream task to the target-language domain. For ***Sensibleness***, we see that the inclusion of more translations yields the best results. As such, we can conclude that low-quality MT does not adversely affect performance. We hypothesise this is due to MT being able to correctly translate keywords that indicate context awareness. Since we are only concerned about relevance, the overall sentence may still contain MT errors and be scored highly.

ML The correlation results for the multilingual models are presented in Table 3. For ***Understandability***, we note that, on average, and similar to LANG, the best performance is attained with the minimum amount of translated data (ML-5), with the performance decreasing when more translations are added. Comparing these results with ChatGPT, we observe an improvement in performance, but our encoder models are still generally weaker when using Spearman as a metric. For ***Sensibleness***, decreasing the amount of data reduces the performance of the model. However, we note a decrease in performance when including the full amount of translated data (ML-100). This may be due to the inclusion of the worst translations – typically hallucinations – which is compounded by training on all languages. Unlike in Understandability, here we see that ChatGPT still outperforms the best encoder model in terms of Pearson correlation.

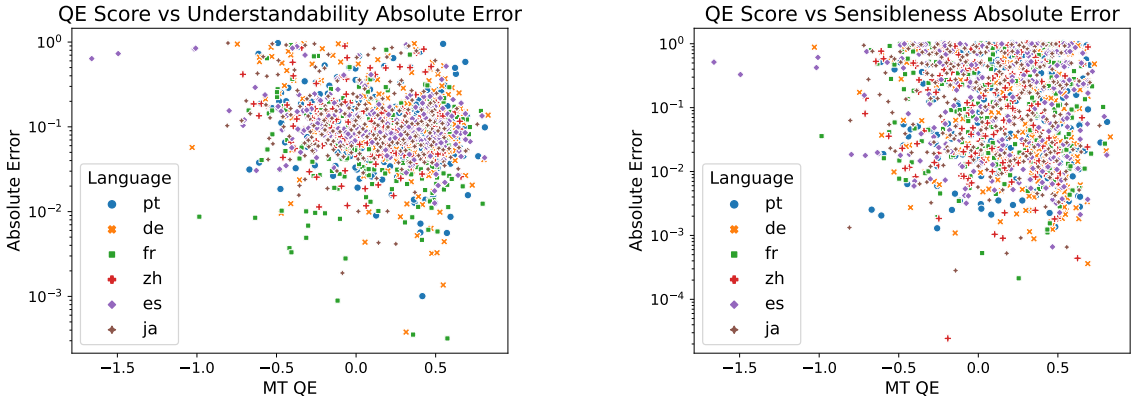
5.3 Effect of low-quality translation during prediction

One might ask if a low-quality translation can induce the submetrics to output a different score. Intuitively, we hypothesise each model will attribute different scores in the face of low quality translations. More specifically, given the results presented in previous sections, we expect the test prediction error to be:

- **Negatively correlated with the MT QE scores for VSP.** We know this model is highly sensitive to low quality translations, since MT errors frequently affect the fluency of the response (as identified in previous sections);
- **Weakly correlated for the NSP model.** The model showed robustness when including more translations during training, with performance decreasing only when we included all translations (ML-100) during training.

In order to evaluate these assumptions, the correlation plots of the MT QE z-scores (obtained independently for each language) against the submetric absolute error using the best ML models (ML-5 for VSP and ML-75 for NSP) for the test set are presented in Figure 4.

For the ***Understandability*** subquality, we note that there is a slight negative correlation between the absolute error and the MT QE score. This is also confirmed by a calculated Pearson Correlation value of -0.245. For the ***Sensibleness*** subquality, the relationship between these two measures is less obvious. For instance, we note that, unlike for Understandability, maximum deviations



(a) Understandability scatter plot.

(b) Sensibleness scatter plot.

Figure 4: Scatter plot comparing the test set MBART50 per-language QE z-scores (x-axis) versus the per sample Absolute Prediction Error (y-axis in log scale) for Understandability and Sensibleness subqualities.

are spread evenly across the QE scale, which points to the model erroneously predicting Sensibleness irrespective of the translation quality. Conversely, we also note a higher density of accurate predictions with lower QE scores. These results, paired with the calculated Pearson Correlation value of -0.129, confirm our hypothesis that the NSP model is more agnostic of MT quality than VSP.

CTX: Também me apercebi desta questão. E a automatização dos processos do escritório é essencial. **RES:** Sim, fazer tudo manualmente demora demasiado.	
EN-VSP: .394	**EN-NSP:** .824
ML-VSP: 1.00	**ML-NSP:** 1.00
Unders.: 1.00	**Sensibl:** 0.00
CTX: Ja, ich leite die Jungs am Kai. **RES:** Wow, das klingt nach einem fantastischen Job, de du da bekommen hast.	
EN-VSP: .963	**EN-NSP:** .315
ML-VSP: .941	**ML-NSP:** .981
Unders.: 1.00	**Sensibl:** 1.00

Table 4: Examples of subquality predictions from the test set.

5.4 Example test predictions

We present representative examples of our best ML models' prediction (ML 5/75) in Table 4. In the first example, the baseline English model fails to appropriately identify the understandability of the response. In the second example, we see that the multilingual model is able to correctly identify that the response takes into account the job presented in the context (manager) by complimenting it ("fantastic job"), which the EN model failed to identify.

6 Conclusions

This paper explored the use of cross-lingual knowledge transfer for the novel task of automatic multilingual dialogue evaluation. We evaluated different strategies for this task, including zero-shot inference, MAD-X and Machine Translation augmentation. Empirically we showed that the naive approach of leveraging MT for augmentation is insufficient to outperform the baseline of English finetuning with a multilingual encoder-based LM, let alone a strong LLM. Instead, by filtering out low quality translations, we were able to reduce the gap of performance on ChatGPT, outperforming it on select correlation metrics. Experimental results showed that we obtain the best performance when training encoder models with the following proportions of MT-QE: 5% for Understandability and 75% for Sensibleness.

One could argue the notion of quality is intrinsically related to cultural norms. For instance, Japanese speakers may prefer a polite conversation, whereas German speakers might prefer a more direct interaction. A future research direction is to evaluate generative model responses in different languages using annotators exposed to the culture associated with a given language. In addition to ensuring the evaluation of the response meets the criteria of "quality" in different cultures, it would also allow for a qualitative analysis of the differences in the notion of quality between languages.

Limitations

Perhaps the main limitation of this work is the restricted amount of languages studied. Ideally, we would have used a more comprehensible set of languages, including low-resource ones, to evaluate the consistency of the conclusions drawn from the experiments.

Another limitation is the focus on a single open-domain dialogue dataset. Dialogue evaluation metrics are known to correlate poorly when evaluated on unseen datasets (Yeh et al., 2021). As such, it is not certain that the observations presented in this work would hold for other datasets, or even different annotations (Mehri et al., 2022).

Finally, the pretrained encoder, MT and QE models used in this work are not fully representative of all available models. We acknowledge that the optimal amount of filtering is likely to be different, depending on the combination of models used.

Ethics Statement

This work leverages dialogues and annotations developed exclusively by English-speakers. This introduces an English-centric bias with respect to the notion of quality (and subqualities) in dialogues. Although not evaluated in depth in this work, there could be a chance that the models erroneously yield lower scores to responses not conforming to English notions of quality responses.

The original dialogue dataset and generated responses were checked for personally identifiable information or offensive content by the original authors. Although highly unlikely, we acknowledge the translations may contain offensive content resulting from decoding.

The post-editing conducted in this work used a crowdsourcing platform that awarded users a fair wage according to their location.

Acknowledgements

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Responsible.AI), and by national funds through *Fundação para a Ciência e a Tecnologia* (FCT) with references PRT/BD/152198/2021 and UIDB/50021/2020, and by the P2020 program MAIA (LISBOA-01-0247-FEDER-045909).

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Dimitra Anastasiou, Anders Ruge, Radu Ion, Svetlana Segărceanu, George Suciu, Olivier Pedretti, Patrick Gratz, and Hoorieh Afkari. 2022. A machine translation-powered chatbot for public administration. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 329–330, Ghent, Belgium. European Association for Machine Translation.

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.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for qa using translation as data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12583–12591.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wei Hu, Zhiyuan Li, and Dingli Yu. 2020. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An

empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, W.K.F. Ngan, Spencer Poff, Naman Goyal, Arthur D. Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv*, abs/2208.03188.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and

Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, João Sedoc, L. F. D'Haro, Rafael E. Banchs, and Alexander I. Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *ArXiv*, abs/2111.02110.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Pengfei Zhang, Xiaohui Hu, Kaidong Yu, Jian Wang, Song Han, Cao Liu, and Chunyang Yuan. 2022a. MME-CRS: Multi-Metric Evaluation Based on Correlation Re-Scaling for Evaluating Open-Domain Dialogue. *arXiv preprint arXiv:2206.09403*.

Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022b. Mdia: A benchmark for multilingual dialogue generation in 46 languages.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

A Training setup and Hyperparamters

We used the XLM-R Large encoder model downloaded from HuggingFace [4] for all experiments. A token representing the speaker was added for each turn, and a history length of 3 turns was used. We applied a regression head consisting of a 2-layer MLP with a hidden size of 1024 and a hyperbolic tangent function as activation for prediction. All parameters were trained/finetuned using Adam optimizer (Kingma and Ba, 2015).

The task adapters were trained using the recipe from Mendonca et al. (2022), using a learning rate of 1e-4 and training for 10 epochs, with a batch size of 32. We used the existing language adapters from AdapterHub whenever possible (EN, ZH, JA) and trained the remaining using the AdapterHub's MLM recipe [5] on Wikipedia data[6]. The fully finetuned models used a learning rate of 3e-6 and were trained for 3 epochs using a batch size of 16. Evaluation was conducted every 1,000 steps for the smaller training sets and 10,000 steps for the larger ones (75% and 100 %). The best performing model on the evaluation set was selected for testing.

For the dialogue data preprocessing we used spaCy [7] and the corresponding core language models. For the translations we used `facebook/mbart-large-50-one-to-many-mmt` from HuggingFace. Batch size was set to 16 and decoding was conducted using beam search, with the number of beams set to 4.

We used a single Quadro RTX 6000 24GB GPU for all experiments.

B Additional Results

Table 5 presents the monolingual model results for the experiments of Section 5. Due to time and computational constraints, we only conduct these experiments using a single seed.

[4] huggingface.co/xlm-roberta-large

[5] github.com/adapter-hub

[6] dumps.wikimedia.org

[7] spacy.io

	EN		PT		DE		FR		ZH		ES		JA		AVG	
	Pr.	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**
Understandability																
0	.347	.192	.381	.176	.353	.184	**.349**	**.106**	**.406**	.251	.372	.210	.268	**.223**	.354	.212
5			**.534**	**.259**	.469	.223	.347	*.095*	.318	**.263**	**.459**	**.223**	.236	.208	**.387**	**.231**
10			.563	.236	**.489**	**.227**	.199	*.102*	.300	.233	.300	.191	**.303**	.206	.357	.218
20			.499	.233	.356	.211	.153	*.082*	.323	.223	.257	.163	.251	.191	.312	.201
50			.433	.214	.418	.185	*.117*	*.017*	.250	.198	.233	.140	.225	.163	.289	.175
75			.186	.189	.306	.158	*.089*	*.026*	.319	.198	.243	.156	*.226*	.185	.245	.169
100			.240	.165	.347	.144	*.082*	*.043*	.248	.206	.191	.109	.216	.146	.239	.155
Sensibleness																
0	.621	.654	.618	.627	.667	.668	.621	.644	.605	.647	.628	.628	.577	.592	.620	.635
5			.615	.636	.687	.657	.632	.628	.618	.629	.599	.631	.538	.553	.616	.626
10			.647	.646	.672	.655	.562	.596	.607	.626	**.635**	.637	.587	.606	.619	.630
20			.639	.644	.680	.679	.627	.640	.620	.633	.615	.634	.582	.595	.626	.638
50			.651	.680	.654	.671	.601	.631	.637	.665	.613	.639	.603	.609	.626	.647
75			.634	.670	.640	.681	**.643**	.664	.629	.673	.615	.639	.608	.635	.627	.656
100			**.671**	**.693**	**.681**	**.698**	.631	**.666**	**.650**	**.688**	.589	**.659**	**.617**	**.633**	**.637**	**.666**

Table 5: Average correlation results for the monolingual models when varying the amount of translated data.