A Context-Aware Annotation Framework for Customer Support Live Chat Machine Translation

Miguel Menezes lmenezes@campus.ul.pt
INESC-ID, Unbabel, University of Lisbon, Lisbon, Portugal
Amin Farajian amin@unbabel.com
Unbabel, Lisbon, Portugal
Helena Moniz helena@unbabel.com
INESC-ID, Unbabel, University of Lisbon, Lisbon, Portugal
João Graça joao@unbabel.com
Unbabel, Lisbon, Portugal

Abstract

To measure context-aware machine translation (MT) systems quality, existing solutions have recommended human annotators to consider the full context of a document. In our work, we revised a well known Machine Translation quality assessment framework, Multidimensional Quality Metrics (MQM), (Lommel et al., 2014) by introducing a set of nine annotation categories that allows to map MT errors to source document contextual phenomenon, for simplicity sake we named such phenomena as **contextual triggers**.

Our analysis shows that the adapted categories set enhanced MQM's potential for MT error identification, being able to cover up to 61% more errors, when compared to traditional non-context core MQM's application. Subsequently, we analysed the severity of these MT "contextual errors", showing that the majority fall under the critical and major levels, further indicating the impact of such errors. Finally, we measured the ability of existing evaluation metrics in detecting the proposed MT "contextual errors". The results have shown that current state-of-the-art metrics fall short in detecting MT errors that are caused by **contextual triggers** on the source document side. With the work developed, we hope to understand how impactful context is for enhancing quality within a MT workflow and draw attention to future integration of the proposed contextual annotation framework into current MQM's core typology.

Keywords

Context-aware error typologies; Machine Translation; Customer Support; Test-Suites; Translation Quality Workflows and Automation; Automatic metrics.

1 Introduction

In past decades, the staggering growth in demand for shared knowledge has led to an increase in translation requests, exceeding human translators' work capacity. In order to accommodate to such request, many enterprises are now integrating MT systems to their workflow that allegedly provide human-like translations in record time. However, despite often claims of

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 286–297
September 4–8, 2023, Macau SAR, China.

human-parity (Xiong et al., 2017), there are plenty of work in the field (Wan et al., 2022; Singh and Singh, 2022) that dispel such allegations, even showing that, under certain circumstances, state-of-the art conventional approaches under-perform and are unable to deal with language nuances, translating words instead of "meanings". Aware of Neural Machine Translation (NMT) limitations, in the last few years, new approaches have been devised to leverage document context for finer-grained MT outputs. Despite sharing similar believes, we suspect that researchers have only now begun to scratch the surface on such complex subject matter, especially when it is not yet clear that context-aware MT systems are indeed able to account for context within a document (Yin et al., 2021). Yet, there is scarce research into document-level MT quality assessment (QA) metrics for more reliable evaluations (Castilho et al., 2020, 2021). Taking into account the present scenario, we propose a framework that deals strictly with context issues instead of relying on more traditional QA metrics regarded as less suitable for document-level NMT assessment. To properly understand the weight of context within a document, we used the previously MQM annotated WMT-Chat-task EN-PT/BR dataset [1], from live chat customer support interactions, creating the perfect test environment for our research, that strives for more equitable and accurate QA MT metrics.

2 State-of-the-Art

It is widely acknowledged that document context is critical for resolving a wide range of translation problems, nevertheless, the sentence-based translation approach remains the most salient characteristic of the prevailing MT paradigm (Post and Junczys-Dowmunt, 2023). This method, in which documents are dismembered in self contained elements (independent sentences) for better translation management, fails in several accounts. First off, the MT system may translate words or phrases based solely on their individual usage, rather than considering their placement in the document as a whole, and second, it largely fails to maintain intersentential relationships within a document (Bawden, 2018). Such behavioral pattern ends up compromising essential textual parameters: cohesion and coherence, giving rise to a warped source text representation. Realizing the limitations of sentence-level MT, in recent years, new proposals have surfaced, encouraging a paradigm shift. Context aware MT models have started to be implemented and designed to leverage contextual information in a document (Zhang et al., 2018; Lopes et al., 2020; Yin et al., 2021), exposing the importance of context in improving MT quality (Nayak et al., 2022), leading to new challenges: how to evaluate the quality of contextual MT models and how to identify if contextual MT models are actually using context?

2.1 Source Contextual Phenomena and Contextual MT Errors Identification.

It can be challenging to identify context-dependent sentences in a document, as well as to detect MT errors caused by a lack of intersentential context in the source document. The difficulty lies in the fact that the definition of context can be problematic as well as circumscribing what is context in a document. Moreover, MT errors that are linked to contextual phenomenon in a source document are often neglected, since, at first sight, they can only be recognized when juxtaposing source and target documents. This comes to show that, to properly assess quality in an MT output, it is essential to acknowledge the importance of the source document, and realize that a source sentence has the potential to bring about a certain set of MT errors that can be mapped to contextual phenomena. We have defined this phenomena as **contextual triggers**, a phenomenon previously observed by Navrátil et al. (2012), when dealing with methods for syntactic source reordering developed for EN-DE, and whose concept support the core aspect for the devised context MT error annotation.

[1] https://github.com/WMT-Chat-task/data-and-baselines

2.2 Context-Aware Typologies

Contextual mechanisms used for developing state-of-the-art context-aware MT models or used for MT QA have been repeatedly explored and studied, with most researchers focusing on the same well-defined contextual categories subset i) anaphoric pronouns, ii) gender and number agreement, iii) lexical ambiguity, iv) ellipsis, v) terminology, vi) discourse connectives, and vii) deixis (Yin et al., 2021; Post and Junczys-Dowmunt, 2023; Castilho et al., 2021). The aforementioned set of contextual categories make up the general framework of analysed issues widely investigated in the literature (Voita et al., 2019; Yin et al., 2021; Lopes et al., 2020). For our research, we aim at analysing and applying these canonical contextual mechanisms that have been continuously addressed for document-level NMT, furthermore, and since previous categories frameworks were developed with generic domains in mind, thus not completely covering the contextual nuances for user generated content in spontaneous dialogues, we have introduce a set of less explored categories that are particularly relevant for the analysed dataset domain, **live chat customer support solutions**. The categories are: Discourse Markers, Greetings, Multiword-Expressions, Named Entities and Register. In tables 2 and 3, we present the complete description of our annotation framework, coupled with examples. Table 2 reflects the mainstream categories accounted for on document-level QA. Table 3, on the other-hand, shows our set of complementary context-categories that can further enhance the identification of **contextual triggers**.

2.3 Metrics for Context Evaluation

Typologies on context are scarce, not suitable for spontaneous dialogues and user generated content. The same applies to context evaluation metrics, that are affected by lack of context examples. One can then assume that insufficient studies on the context evaluation metrics as well as insufficient training data for contextual MT evaluation have detrimental consequences in MT QA results. This section will cover QA in general and how it has been applied to context.
Currently, MT outputs quality evaluation is performed relying on both automatic evaluation metrics, e.g., COMET (Rei et al., 2020), chrF (Popović, 2015), SacreBLEU (Post, 2018) as well as on human judgments, using, for example, the MQM Framework Typology (Lommel et al., 2014). MQM with its hierarchic error typology framework, easily adapted by users according to particular needs with a total of 100 issue types with various levels of granularity, has not been created to have in mind contextual MT errors, which does not prevent it from being applied to QA of context-aware MT models (Freitag et al., 2021, 2022), leading to unreliable results. We regard current MQM framework as unfit to fully deal with contextual nuances, creating potential biases in document-level NMT QA results.
Moreover, concerning automatic document-level NMT QA, the current practice is to resort to existing pretrained models e.g., BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020) by simply providing several sentences of context to the pretrained model, allowing the pretrained model to use surrounding context. We hypothesize that this technique of leveraging existing sentence-level metrics might not be conducive to robust enough models capable of covering the complete spectrum of contextual errors.

3 Multilingual Virtual Agents for Customer Service (MAIA) Corpus

For our research, we used the MAIA corpus (Farinha et al., 2022), made available for the WMT 2022 Shared Task on Chat Translation, containing genuine bilingual customer support interaction (chat conversation between customer support agents and customers). Such content is planned on-the-fly and written on-line, usually coupled with abbreviations, emoticons, idiomatic expressions and grammatical and typographical errors. We took advantage of this ideal test environment to i) understand how context is conveyed in a document, ii) pinpoint lexical

structures linked to contextual information, iii) create an annotation framework that allows to measure context in a document, with the needed plasticity to be added to more traditional quality measure metrics iv) give the first steps on creating a multilingual test suite with contextual annotations for real customer support data.

Maia Corpus	**EN-PT/BR**
Number of conversations	28
Number of agent segments	509
Number of customer segments	609
Number of total (customer and agent) segments	1168

Table 1: Statistics of the dataset used for context annotation.

4 Contextual Annotation Framework.

Recent context-aware MT models progress calls for developing new evaluation solutions that cover contextual errors. Our framework allows to identify and classify contextual discourse structures linked to MT errors. This section will initially describe the most frequently addressed contextual categories in the literature, followed by our new set of contextual categories found to be relevant for the customer support live chat domain data. Note that the framework was created with the possibility to be accustomed to other domains.

4.1 Building a Context-Aware Typology

To devise a contextual framework, we built on previous works, such as the Document-Level Machine Translation Evaluation (DELA) by Castilho et al. (2021) that introduces several meaningful contextual related issues, *e.g.*, Agreement; Ellipsis; Gender Agreement; Lexical Ambiguity; Terminology; and Number. Using a corpora-based analyses approach of an ecological dataset, we aimed to explore the standard categories proposed in the literature. Consequently, we extended our analysis to consider less explored contextual categories, such as, **Discourse Markers, Greetings, Multiword Expressions, Named Entities and Register**, which have a significant impact in the chat domain. The identification of the contextual issues entailed an annotation step where the **contextual triggers** were identified and categorized. To the best of our knowledge, our research is the first to focus on contextual issues for MT for the customer support chat domain. Next, we will introduce all the contextual categories that compose our framework, starting with the more explored-canonical categories, followed by the new proposed categories, see Tables 2 and 3.

4.2 The Annotation Process

The annotation process was performed by a Portuguese annotator with a background in translation and with previous experience in contextual issues annotation. Concerning the test sets, we used the official submissions of the WMT-Chat-2022 shared task for the EN-PT/BR language pairs, translated by two MT systems: Baseline and Unbabel-IST. Note that, the dataset used came already with a prior MQM non-contextual annotation performed for the WMT 2022 Chat Shared Task. Both MT systems are based on the large multilingual pre-trained models. The Baseline model, uses a vanilla M2M-100 model, Fan et al. (2021), while the Unbabel-IST model uses a fine-tuned version of mBART50, Liu et al. (2020). For the fine-tuning data, it uses the in-domain parallel validation set provided by the shared task organizers and a generic parallel corpus. For our analysis, only the sentences requiring context with a MT issue/error have been considered. For those, the annotator performed as follows: i) identified the **contextual trigger** that caused the MT error, ii) categorized it, providing a translation, iii) identified

Category	Example and Explanation
Agreement: Targets gender and number agreements.	*Source:* Por quanto tempo vou poder ficar afastada? *Target:* How long will I be able to stay away? *Source:* While your account is on pause, you will not be billed for a new month subscription. *Target:* Enquanto sua conta estiver em pausa, você não será cobrado/a para um novo mês de assinatura.
	Explanation: Gender agreement: masculine cobrado/ feminine cobrada beyond the sentence level. In the example, only by accessing previous information (context afastada) we are able to understand that we need the feminine translation cobrado/a.
Lexical ambiguity: Refers to the polysemy of words in distinct contexts.	*Source:* Thanks so much for your interest in partnering with us *Target:* Obrigado por seu interesse em colaborar connosco! *Source:* Someone on our Corporate team will reach out *Target:* Alguém em nossa equipe corporativa chegará. (Glosa: will arrive).
	Explanation: The translation of "reach out" requires information that lies beyond the sentence, assuming a complete different meaning from arriving. Correct Translation: Alguém em nossa equipe corporativa "**entrará em contacto**". Glosa: will contact you
Ellipsis: Refers to omission of word(s) within a sentence. Syntactically, the linguistic information is recovered.	*Source:* It looks like this inquiry requires further investigation, and we'll need to log into a few different systems. *Target:* Parece que esta pesquisa requer mais investigação e precisaremos de entrar em alguns sistemas diferentes. *Source:* Quando [-] forem consultar a principal questão é sobre os créditos não expirarem mais *Target:* When they go to consult, the main question is about the credits do not expire more
	Explanation: the elliptical pronoun **[-]**, wrongly translated as **they**, is only recovered accessing previous sentences: "**we**'ll need to log into a few different systems". Correct translation: When **you** go to consult (...).
Terminology: Targets terms that constitute a set of vocabulary within a specialized field of knowledge.	*Source:* On your phone or tablet, open the #PRS_ORG# app. *Target:* No seu telefone ou tablet, abra a aplicação #PRS_ORG# . *Source:* At the top right, tap More. *Target:* Na parte superior direita, clique em Mais. *Source:* Tap history. *Target:* Tap história.
	Explanation: Contextually, the word "history" is a term and should be translated as **histórico**. In this case, the MT does not recognizes "history" as a term.

Table 2: Conventionally context categories used for annotation.

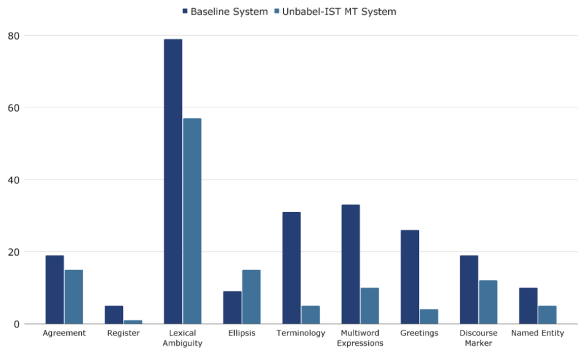


Figure 1: Contextual categories error distribution for each MT model

the turn that serves as anchor to disambiguate the issue, and iv) attributed a level of severity for each issue.

5 Results

In this section we analysed the errors and characterized them according to the Context Aware Typologies Framework that we developed, providing an error and severity analysis, whilst, simultaneously, contrasting our annotation with the MQM's non-contextual annotation performed previously for the WMT 2002 Chat Shared Task.

5.1 Context Dependent Segments

From the dataset with 1168 sentences, we identified 197 sentences (17% of the dataset) with MT errors that can be mapped to **contextual triggers** for the Baseline model, and 123 sentences for the Unbabel-IST model (10% of the dataset).

5.2 Contextual Categories Distribution

Figure 1 displays the contextual categories distribution linked to the MT errors in our dataset. As seen in Figure 1, the most prevalent MT errors are induced by *Lexical Ambiguities* in the source document, 76 MT errors for the baseline MT system and 56 MT errors for the Unbabel-IST model. Taking into account the overall MT errors linked to our contextual categories per MT model, we observe that the presence of lexical ambiguities in the source document accounts for 34% of the overall contextual MT errors for the baseline MT system, and 45.50% for the Unbabel-IST model. Note that, since the percentages were calculated taking into account the MT overall contextual errors outputs **for each MT model** (the baseline MT system outputted 231 contextual errors, the Unbabel-IST model 124), the percentage values reflect the weight that each category has within those subsets (the overall contextual errors for each system). Concerning the category *Terminology*, the Baseline showed 31 MT errors, accounting 13% of the overall MT contextual error, and 5 MT errors for the Unbabel-IST system, accounting 4% of the overall MT contextual error. This difference can be explained by the fact that the second model was fine-tuned with the in-domain data and was specialized to this domain, and not necessarily by its ability in handling the contextual terminology errors.

For the category *Multiword Expressions*, the Baseline model reports 33 MT errors, 14% of the total contextual errors for this model, whilst the Unbabel-IST system reports, 10 MT errors, accounting 9%. *Agreement* is a very present error category within the analysed dataset. This category is particularly relevant, since it deals with gender agreement, and it is considered a

Category	Example and Explanation
Discourse Markers: Fillers or other words that are used to indicate dialogue interactions. Different discourse markers convey different meanings for the fluidity of a dialogue.	*Source:* Thank you please try the following steps: *Target:* Obrigado, por favor, tente os seguintes passos: *Source:* Delete cache, restart your device *Target:* Delete cache, reiniciar o seu dispositivo *Source:* Tá bom *Target:* It is good
	Explanation: The expression "Tá bom" should have been translated as an acknowledgment discourse marker, such as "**ok**", instead it is literally translated as "it is good".
Greetings: Conventionalized expressions used as part of our daily lives when greeting, well-wishing and leaving a conversation. These structures are dependent on the degree of politeness and cultural awareness.	*Source:* Bom dia. *Target:* Good day. *Source:* Gostaria de saber melhor como funciona os créditos. *Target:* I would like to know better how the credits work.
	Explanation: The expression "Bom dia", can be translated in EN as "Good day" meaning "it is a good day", but it should have been translated as a greeting "Good morning", "Hello". Since greetings are culturally and language dependent, they are are negatively influenced when contextual information is scarce.
Multiword-expressions: Compounded units, e.g., phrasal-verbs, they act as a single unit. These structures can either be solved within a sentence or require contextual information to be disambiguate.	*Source:* Cancelei meu plano mas mesmo assim me cobraram. *Target:* I cancelled my plan but still they charged me. *Source:* Thank you for reaching #PRS_ORG#! *Target:* Obrigado por entrar em contacto com #PRS_ORG#! *Source:* Let me check on that for you. *Target:* Deixe-me verificar isso para você. *Source:* Please hold while I pull up your account. *Target:* Por favor, mantenha enquanto eu retirei sua conta.
	Explanation: The Multiword-expression "pull up" was translated as "retirar" (to withdraw), but in the specific context the correct translation would be: enquanto **acesso** à tua conta (glosa: whilst I access you account).
Named Entity (NE): Linguistic structures which refers to, e.g., a book title , a person's name, an address, a credit card number.	*Source:* Boa tarde, não consigo comprar livros com nenhum cartão de crédito apenas com cartão de oferta. *Target:* Good afternoon, I can't buy books with no credit card only with offer card. *Source:* O último foi hoje, à pouco e chama-se A única mulher. *Target:* The last was today, shortly, and it is called the only woman.
	Explanation: The NE, a book title ("A única mulher"), is not identified within the sentence and should not have been translated, since the user is looking for the book in Portuguese, but the original book´s name was translated.
Register: Degrees of politeness where speakers adapt their discourse according to the audience.	*Source:* How can I help you today? *Target:* Como posso te ajudar hoje?
	Explanation: In the example, "help you / **ajudar-te**" is not appropriate, since it is using a very informal second person singular. The correct translation would be: Como posso **ajudá-lo/la**?, a third person singular.

Table 3: New set of contextual categories triggers.

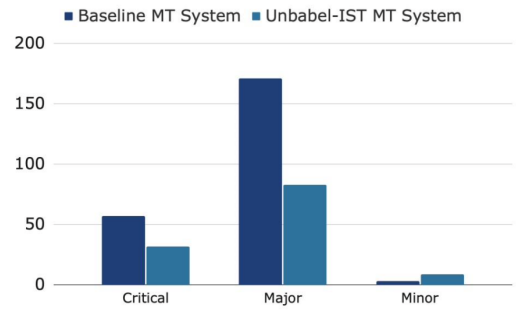


Figure 2: Contextual error severities for each MT model

critical error. For this case, the Baseline shows 19 gender agreement MT errors, about 8% of the MT contextual errors for this model, whilst the Unbabel-IST model shows 15 MT gender agreement errors, about 12% of the complete set of MT contextual errors, All things considered, although we see that the Unbabel-IST model produces significantly fewer contextual errors, this can be simply due to the domain-adaptation effect and not necessarily in its capability to deal with the contextual phenomena.

5.3 Categories (Not) Covered by the Core MQM Framework

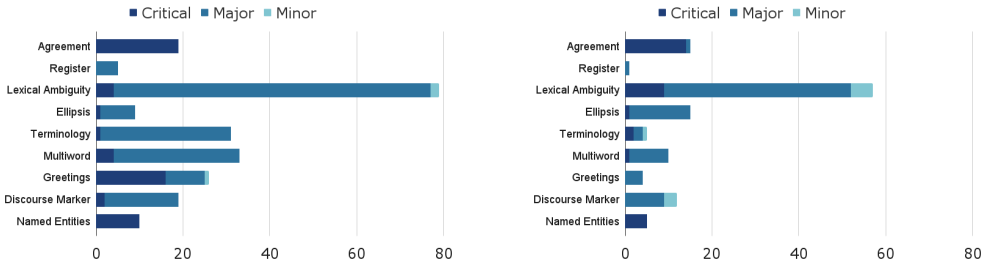
In our research, we have noticed that core MQM typology used for the WMT-2022 chat shared task moderately identifies some contextual issues, in part because annotators were instructed to, if possible, account for some dependencies within the dataset. Nevertheless, 36.1%, for the Baseline and 42% for Unbabel-IST of the contextual issues annotated by the Context-Aware Typology were not considered during the WMT-2022 chat shared task MQM annotation. Concerning the contextual issues identified by the MQM, they were tagged as **Mistranslations** in most cases, without specifying the underlying cause, e.g., an absence of context at a sentence level. As such, **Multiword Expressions, Discourse Markers, Lexical Ambiguities and Greeting** errors, according to MQM analysis results, were annotated as **Mistranslations**. Moreover, these errors fall for the most part within the critical and major error severity, compromising customer/agent communication fluidity.

5.4 Contextual Categories Distribution Severities

As Figure 2 displays, most contextual issues fall under the severity Critical, 24.6% for the Baseline, 25.8% for the Unbabel-IST MT model; and Major, 74% for the Baseline, 66% for the Unbabel-IST MT model. These errors severely compromise understanding and communication, impacting customer support reliability. Concerning the Minor severity, those values present strictly residual numbers, reinforcing the importance of contextual issues.

5.5 Contextual Error Severities by Categories and MT Model

As seen from the charts in Figure 3, there is a considerable difference between models concerning the total of contextual issues. Nevertheless, there are similar patterns regarding some categories. According to the tables, lexical ambiguity issues, considered a Major error, are common and make a considerable amount of the issues for both models. The category *Agreement* shows a sizable value for both models, being considered for most cases a Critical issue. *Terminology*, *Multiword Expressions* and *Discourse Markers* are categories particularly interesting to observe, due to their disparity between the baseline and Unbabel-IST model. This difference validates the hypothesis that models trained with in-domain datasets are more robust,



(a) baseline (b) Unbabel-IST

Figure 3: Distribution of error categories and their severities of a) the baseline MT system, and b) the Unbabel-IST system.

outweighing some contextual issues. Nevertheless, despite showing significant quality output improvements, robust models still fall short in detecting contextual nuances, substantiating, and validating future research in document-level MT models.

6 Automatic Metrics of MT Evaluation and Contextual Errors

Measuring the ability of the current state-of-the-art MT evaluation metrics in detecting the contextual errors is the first crucial step for developing new automated quality evaluation solutions for the MT systems using our proposed typology. Hence, we measured the correlation of these metrics with the MQM annotations of the MAIA test-set. To have a reliable term of comparison, in addition to the contextual annotations, we also measured the correlation of the metrics on the original MQM annotations based on the existing framework.

For the metrics, we used COMET (Rei et al., 2020) that is trained to predict the human translation quality judgments of the MT outputs. It evaluates the translations in isolation without considering their contexts at all. Very recently, Vernikos et al. (2022) introduced an extension of this metric (i.e., Doc-COMET) that incorporates context when evaluating the MT outputs. Vernikos et al. (2022) show that Doc-COMET obtains a higher system-level Pearson correlation with human judgments compared to its original sentence-level counterpart on TED talks and News domains for En-DE, En-RU, and ZH-EN language pairs.

Since the system-level analysis does not provide detailed insights on the ability of the metrics in capturing the contextual errors, we focused our analysis on the sentence-level correlation of the metrics with human judgments on the MAIA dataset. Given that our framework is tailored for the contextual errors only, for our analysis we concentrated on the samples that contain at least one contextual error in the output of the MT system. We also made sure that errors that do not have a contextual background have no reflection on the automatic metrics results. To this aim, and to not lose the context, we first obtained the scores of all the sentences of the test-set with each metric, and then used only the segments with contextual errors, 197 sentences, for the baseline model, and 123 sentences for the Unbabel-IST model.

Table 4 shows the sentence-level Pearson correlations of the two metrics for both MT systems. As the results suggest, both COMET and Doc-COMET have a lower correlation with the MQM scores of our annotation framework. This, however, is expected mainly because the COMET models were trained on the data annotated with the existing MQM annotation framework. Moreover, we clearly see that there is no reliable correlation between DocCOMET and the human judgments of both frameworks on the sentence level. This can be justified by the fact that the COMET models were not trained on any document-level annotations, hence they

Metric	Baseline	Unbabel-IST
Correlation with the existing error annotation framework		
COMET	0.35	0.35
Doc-COMET	0.13	0.07
Correlation with our contextual error annotation framework		
COMET	0.25	0.06
Doc-COMET	-0.07	-0.19

Table 4: Sentence-level Pearson correlation of COMET and Doc- COMET metrics with MQM annotations on a subset of the test-set that contains at least one contextual error. The annotations are done with the existing framework and our new contextual errors framework.

cannot detect contextual errors accurately.
These findings show that in order to measure the quality of the MT systems on the contextual errors, new datasets, metrics and tools need to be developed that not only cover the existing sentence-level errors, but also can cover the contextual errors that none of the current resources cover, and usually are categorized as severe errors (i.e., either critical or major).

7 Conclusion

With our research, we have shown the significance of context for the MT. Similarly, we exposed the inadequacy in conventional QA metrics for reliable qualitative assessments, since current QA models and frameworks show to be weak and deceptive as they have not been created to have in mind contextual MT errors. We have displayed first attempts in overcoming QA models and frameworks shortcomings in the form of contextual errors test-suites, but also those are scarce in terms of contextual typologies coverage and focus on common analysed domains. We instead propose an alternative contextual framework for document level MT QA, covering a relatively untapped domain in terms of contextual errors analysis. Our framework shows significant gains of an average of 61% more contextual errors coverage than more conventional QA metrics, highlighting the fact that most of such contextual errors are deemed as critical and major, thus strengthening our beliefs that the field of QA for context aware MT is far from being effectively dealt with, on the one hand, and that contextual error severely compromise MT outputs, on the other hand.

8 Future Work

We are well aware of several research limitations in our work, as such, we intend to address these in future work. We aim to apply our framework to different domains and different language pairs, for that, we plan to resort to a team of expert annotators, allowing us to extensively put to test and validate our framework.

Acknowledgments

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), under project UIDB/50021/2020; by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; by the project Multilingual AI Agents Assistants (MAIA), contract number 045909. This work was also supported by the FCT PhD grant with the reference 2022.12091.BD.

References

Bawden, R. (2018). *Going beyond the sentence: Contextual machine translation of dialogue*. PhD thesis, Université Paris-Saclay (ComUE).

Castilho, S., Cavalheiro Camargo, J. L., Menezes, M., and Way, A. (2021). DELA corpus - a document-level corpus annotated with context-related issues. In *Proc. of the Sixth Conference of WMT*, pages 566–577.

Castilho, S., Popović, M., and Way, A. (2020). On context span needed for machine translation evaluation. In *Proc. of the Twelfth LREC*, pages 3735–3742, Marseille, France. European Language Resources Association.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2021). Beyond english-centric multilingual machine translation. 22(1).

Farinha, A. C., Farajian, M. A., Buchicchio, M., Fernandes, P., C. de Souza, J. G., Moniz, H., and Martins, A. F. T. (2022). Findings of the WMT 2022 shared task on chat translation. In *Proc. of the Seventh Conference on Machine Translation*, pages 724–743, Abu Dhabi, UAE.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. (2022). Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proc. of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, UAE.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. (2021). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proc. of the Sixth Conference of WMT*, pages 733–774.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the ACL*, 8:726–742.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Lopes, A., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. T. (2020). Document-level neural MT: A systematic comparison. In *Proc. of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Navrátil, J., Visweswariah, K., and Ramanathan, A. (2012). A comparison of syntactic re-ordering methods for english-german machine translation. In *Proc. of COLING 2012*, pages 2043–2058.

Nayak, P., Haque, R., Kelleher, J. D., and Way, A. (2022). Investigating contextual influence in document-level translation. *Information*, 13(5):249.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proc. of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proc. of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium.

Post, M. and Junczys-Dowmunt, M. (2023). Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv:2304.12959*.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Singh, S. M. and Singh, T. D. (2022). Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Systems with Applications*, 209:118187.

Vernikos, G., Thompson, B., Mathur, P., and Federico, M. (2022). Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proc. of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, UAE.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proc. of the 57th Annual Meeting of the ACL*, pages 1198–1212, Florence, Italy.

Wan, Y., Yang, B., Wong, D. F., Chao, L. S., Yao, L., Zhang, H., and Chen, B. (2022). Challenges of neural machine translation for short texts. *Computational Linguistics*, 48(2).

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2017). Achieving human parity in conversational speech recognition.

Yin, K., Fernandes, P., Pruthi, D., Chaudhary, A., Martins, A. F., and Neubig, G. (2021). Do context-aware translation models pay the right attention? *arXiv preprint arXiv:2105.06977*.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018). Improving the transformer translation model with document-level context. *arXiv preprint arXiv:1810.03581*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.