Quality Fit for Purpose: Building Business Critical Errors Test Suites

Mariana Cabeça[†γι*] **Marianna Buchicchio**[‡ι*] **Madalena Gonçalves**[οι*]
Christine Maroti[‡ι*] **João Godinho**[οι*] **Pedro Coelho**[οι*] **Helena Moniz**[αγι] **Alon Lavie** [οι]
[ι]Unbabel [γ]INESC-ID
[†] marianacabeca14@inesc-id.pt
[ο]{firstname.lastname}@unbabel.com
[‡] {firstname}@unbabel.com
[α] helena.moniz@inesc-id.pt

Abstract

This paper illustrates a new methodology based on Test Suites (Avramidis et al., 2018) with focus on Business Critical Errors (BCEs) (Stewart et al., 2022) to evaluate the output of Machine Translation (MT) and Quality Estimation (QE) systems. We demonstrate the value of relying on semi-automatic evaluation done through scalable BCE-focused Test Suites to monitor both MT and QE systems' performance for 8 language pairs (LPs) and a total of 4 error categories. This approach allows us to not only track the impact of new features and implementations in a real business environment, but also to identify strengths and weaknesses in models regarding different error types, and subsequently know what to improve henceforth.

1 Introduction

Unbabel's Language Operations platform blends advanced artificial intelligence with humans in the loop for fast, efficient and high-quality translations that get smarter over time. The company combines Machine Translation with Human Post-Edition performed by experienced post-editors to translate a variety of content, ranging from Customer Support to Marketing. MT and Quality Evaluation are at the core of Unbabel's business, as the main focus is to provide high-quality translations regardless of the use case or content type. Both MT and QE systems have been continuously improving and overcoming existing limitations throughout the years. As a result, the need to evaluate their outputs' accuracy and overall performance in error detection grows along with this development process, especially in a business environment where the need to deliver high quality translations without critical errors is paramount.

The evaluation of MT outputs can be generally done by following either manual quality assessment procedures with error annotations (such as the Multidimensional Quality Metrics (MQM) Framework (Lommel et al., 2014)), or automatically by relying on metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and COMET (Rei et al., 2020).

In the same way that MT evaluation and complementary metrics are crucial to achieve outputs with better quality, so is the need to evaluate the precision and accuracy of QE systems. To this end, QE systems are oftentimes evaluated against gold annotated data by the Pearson correlation score (e.g., Fonseca et al. (2019)) and by computing the Matthews correlation coefficient (MCC, (Matthews, 1975)).

The main focus of this paper is to overcome the shortcomings of both manual and automatic MT and QE evaluation methodologies in a 'real-life' business environment. We are able to achieve this through a semi-automatic approach that relies on MT Test Suites (Avramidis et al., 2018) in a production setting. The MT Test Suites proposed here follow the concept of BCEs (Stewart et al., 2022) and consist of proprietary corpora with MQM-annotated data (Lommel et al., 2014). With this in mind, we demonstrate how MT Test Suites can be leveraged to provide a semi-automatic method of MT evaluation and how they can be a good compromise between manual and automated metrics by taking into account errors that are harmful in

*These authors contributed equally.

a business environment.

In this paper, we seek to address the following:

1. How can we improve MT evaluation by relying on Test Suites focused on critical errors in a business environment?

2. How can we evaluate QE systems and appraise their rigor in error detection tasks?

For this purpose, we present large-scale and fine-grained MT Test Suites for 8 LPs with English as source language for all possible combinations. As we base our approach on the concept of BCEs (Stewart et al., 2022), the MT Test Suites proposed here will be called BCE Test Suites.

2 Related Work

The MT field witnessed a breakthrough in the quality of translations with the rise of Neural MT (NMT). As such, the need for evaluating the performance of different systems increased concurrently. There are two major types of approaches when evaluating MT systems: manual and automated metrics.

Regarding manual metrics, one distinctive method has been widely adopted in an attempt to standardize the evaluation process: the MQM Framework (Lommel et al., 2014), which provides a hierarchical categorization of issue types and dependencies regarding errors in translation outputs. Each error is annotated by human annotators with a precise issue type, along with the level of severity that affects the target text and its perceived quality. There are three severity levels an error can be classified as: minor, major, and critical. However, it is important to stress the difference between critical errors from a *linguistic quality* perspective and errors related to the *perceived quality* of the translation. While critical linguistic errors severely impact the grammaticality of the text, errors that disturb the perceived quality of a translation are considered as Business Critical Errors (BCEs) (Stewart et al., 2022). This is due to the fact that they not only include errors that are considered linguistically critical, but also errors that may cause additional damage in a customer-focused environment.

In addition to manual processes of MT evaluation, automatic metrics have also been commonly adopted in the industry to assess the MT outputs' quality along with the MT systems overall performance. Two examples of these metrics, among the most commonly used ones, are: BLEU (Papineni et al., 2002), that estimates a translation's quality value by solely relying on Precision, and COMET (Rei et al., 2020) — a widely-used recent metric developed by Unbabel. COMET is a neural framework that allows multilingual MT evaluation and that highly correlates with human judgement.

Another way to measure systems' performance is by automatically estimating the quality of the translation without access to a reference. Specia et al. (2009) and Kepler et al. (2019) are able to achieve this through the use of QE metrics. While Specia et al. (2009) estimate quality by relying on a continuous score, Kepler et al. (2019) took a step forward for QE tools and created a new open source framework named OPENKIWI[1]. OPENKIWI, created by Unbabel, achieved state-of-the-art results in word-level QE at the time. Following that, Unbabel also won the WMT 2022 Shared Task on Quality Estimation (Zerva et al., 2022) with an extension of the COMET (Rei et al., 2020) framework called COMETKIWI (Rei et al., 2022), which merges the benefits of COMET's multilingual training features with OPENKIWI's predictor-estimator architecture.

Despite all the advancements of the automatic evaluation approaches, the existing solutions fail, to some extent, to detect BCEs. In order to relax this issue, there are several approaches to data augmentation, such as AugLy (Papakipos and Bitton, 2022) and, more recently, Alves et al. (2022) who proposed a new Sentence-level Multilingual AUGmentation (SMAUG) framework that generates critical errors in translations in order to improve robustness of state-of-the-art MT metrics.

Although both evaluation methods allow for a performance comparison of MT systems, each one shows different advantages and constraints. While automated metrics are unable to provide information about translation error types, they provide a reproducible generic score of correctness (Macketanz et al., 2022) in a time- and cost-efficient manner. On the other hand, manual evaluation is time-consuming and less scalable than automatic methods as it consists of plain human judgement. Nonetheless, manual evaluation is able to provide evaluations that are much more fine-grained and sensitive to nuanced errors. With this in mind and in an attempt to achieve a more detailed qualitative analysis on performance evaluation, a semi-

[1] https://unbabel.github.io/OpenKiwi/

automatic approach that relies on previously revised test sentences to evaluate performance of MT systems was developed in order to merge the advantages of both methods. These test sentences are specifically assembled to obtain corpora of controlled examples, i.e., to obtain Test Suites. The chosen examples in Test Suites are referred to as the gold-standard data and are used for diagnostic evaluation of MT systems. Depending on the type of evaluation desired, Test Suites can be adapted to fit different purposes. As such, they can focus on more specific linguistic phenomena (Guillou et al., 2018) or on generic system's evaluation, as well as being created upon fabricated examples or representative content translated by the MT systems. Thus, their construction is required to follow a linguistically motivated approach, which allows them to be used for comparative analysis between systems (Macketanz et al., 2022; Avramidis et al., 2018).

In sum, by combining manual evaluation with automated metrics, it is possible to obtain values that are much more precise and accurate at describing systems' performance and at identifying the most problematic structures.

3 Methodology

As a means of measuring translation quality, Unbabel performs MQM annotations by using a proprietary MQM-compliant typology adapted from the original MQM proposed by Lommel (2014). Annotations are performed by Unbabel's Professional Community of Annotators, composed of professional translators and linguists with significant experience in linguistic annotations and the detection of translation errors. The result of this process is not only an MQM score that indicates the quality of a given translation, but also annotated data with precise information about the types of errors and the associated severities that occur in MT outputs. Besides this, Unbabel developed the concept of BCE (Stewart et al., 2022), a subset of error categories that can have direct business implications for customers and that would otherwise render a translation 'unfit', regardless of perceived linguistic quality. With MQM annotations we are able to identify the relevant BCEs produced by MT systems and we use them as the basis to build BCE Test Suites. The BCE Test Suites proposed here consist of a total of 8 LPs (Table 1), 4 categories of translation errors with high impact on customers according to the definition of Business Critical Errors (Stewart et al., 2022):

1. *Agreement*: two or more words do not agree in case, number, gender or other morphological feature;
2. *Wrong Named Entity*: any type of mistranslation that affects a Named Entity;
3. *Register*: when the text uses the wrong register (i.e., the level of formality required) for instance expressions, pronouns and verbs;
4. *Untranslated*: a word or a phrase that should have been translated was left untranslated.

With this, we produced a total of 11,481 test sentences, in which each test sentence represents one single error type. For each one of the 4 categories that compose the BCE Test Suites, we aimed at a minimum of 50 test sentences.

LP	Number of Segments
en–ru	2908
en–es–latam	2102
en–es	1820
en–fr	1749
en–it	1180
en–de	805
en–pt-br	702
en–zh-cn	215

Table 1: Total number of Test Suites segments per LP.

Finally, we followed a similar approach to the one proposed by Avramidis et al. (2018), but, instead of applying regular expressions to the test sentences, we used Unbabel's proprietary corpus of MQM-annotated data of in-house MT systems and provided the gold translation of each error. Furthermore, in order to reach the minimum limit of 50 test sentences per category, we performed critical errors data augmentation by following the approach proposed by Alves et al. (2022) for a targeted set of Named Entities.

The methodological process involved in creating the BCE Test Suites along with the curation step performed by in-house professional translators and linguists allowed Unbabel to overcome two major limitations of publicly available similar work (e.g., Isabelle et al. (2017); Avramidis et al. (2019); Macketanz et al. (2022)). These limitations are as follows: Test Suites usually target

a reduced number of LPs; and the focus of Test Suites is oftentimes on specific linguistic phenomena that may not be representative of 'real-world' MT outputs. We aim to overcome such limitations due to the fact that not only the BCE Test Suites account for 8 different target languages, but also because they consist of content already translated by Unbabel's MT systems, thus providing a suitable evaluation that is representative of systematic core errors.

3.1 Building the BCE Test Suites

As mentioned in Section 3, the BCE Test Suites corpus is built by using source and target pairs previously annotated with Unbabel's proprietary MQM-compliant error typology. After the annotation process, we isolated the BCEs that were relevant for the purpose of the BCE Test Suites. In order to build the corpus, we retained information about the LP, the required register, the source and target texts, the annotated error, the error category (according to MQM) and the related severity[2]. Retaining the information about severities was fundamental as we based our methodology on the BCEs definition and removed unnecessary minor errors.

As stated in Section 3, the minimum number of test sentences per category was set to 50 and there were instances in which we needed to perform data augmentation to reach this target, especially in the case of *Wrong Named Entity*. For this reason, we followed the approach proposed by Alves et al. (2022) and applied the SMAUG Framework to introduce deviations in Named Entities and Numbers for the supported LPs, such as "English–German", "English–Spanish", "English–Spanish-Latam", "English–French", "English–Italian" and "English–Simplified Chinese". Finally, the BCE Test Suites were manually curated by in-house linguists specializing in Translation Studies and Computational Linguistics who reviewed the annotations performed by Unbabel's Professional Community and then provided the gold standard of the annotated errors. The linguists were native speakers or with high proficiency in the LPs taken into account. In order to avoid over-penalizing the evaluation, linguists were also asked to exclude cases in which one error would possibly have multiple solutions of translation. The final number of test sentences per LP can be found in Table 1.

Finally, the BCE Test Suites are stored in a specific data-set management system and the metrics are widely available to the business through a Business Intelligence (BI) platform. Section 3.1.1 and Section 3.1.2 will outline how the resulting metrics are computed and applied to MT and QE evaluation.

3.1.1 BCE Test Suites for Machine Translation

The BCE Test Suites are used as a means of MT model evaluation and are used to test the ability of the models to avoid certain BCEs (Stewart et al., 2022) and also as a regression test set.

At Unbabel, we run frequent and periodic retrainings of our MT models. At the end of the training, the new version of the model is evaluated on several data-sets. One of the extracted metrics is the accuracy on each BCE Test Suite, which is defined by matching the 'gold translation' tokens to the respective ones in the MT output.

3.1.2 BCE Test Suites for Quality Estimation

The BCE Test Suites can also be used to evaluate QE systems on error detection for specific category types. To adapt the BCE Test Suites to the QE setting, we run the QE system on the source and MT containing the targeted error, and check that the QE-predicted tag for the error is 'BAD'. If the error spans multiple tokens, we consider the error detected if QE labels any of the incorrect tokens as 'BAD'. This method only measures error recall for the specific error being targeted, since we do not store information about all of the other errors in the sentence in the BCE Test Suites. The final metric reported is the number of segments for which QE caught the error divided by the total number of segments in the BCE Test Suites.

At Unbabel, we use Business Critical Error recall as an additional signal when evaluating QE systems to be put into production. Pure sentence-level or word-level correlations with gold annotations do not always tell the full story when it comes to evaluating QE for a real business use case.

4 Experimental Setup

The main purpose of the BCE Test Suites is to evaluate the ability of MT and QE systems to avoid or

[2]Business Critical Errors are defined by the relevant MQM error category and the severity attributed by the annotator. Moreover, BCEs are defined according to a certain level of quality to be expected for a precise use case. At Unbabel, we identified 5 different levels of translation quality and the relevant BCEs can be consulted here: https://github.com/Unbabel/EAMT23-BCE-Test-Suites/blob/main/BCEs_and_quality_levels.png

detect certain types of errors that can potentially be harmful to customers, according to the type of content and the level of quality expectations related to it. In this paper, we aim to test and measure the behavior of such systems in a real business scenario, especially in regards to the implementation of new features in customers' MT systems and a new MQM-QE model.

4.1 A subset of BCE Test Suites

As mentioned in Section 3, one of the 4 error categories included in the BCE Test Suites is *Wrong Named Entity*, and, because of its broad definition we decided to divide it into more fine-grained categories, such as: *City*, *Country*, *Currency*, *Date* and *Products and Organizations* (*PRS/ORG*). Furthermore, the focus of the experiments was to test the ability of MT and QE systems to handle certain types of Named Entities, as their mistranslation can be dangerous for customers, so the fine-grained analysis is more informative than the broad category. In order to create a subset of the original BCE Test Suites, we used Unbabel's proprietary Named Entity Recognition System (NER) (Menezes et al., 2022; Mota et al., 2022) to automatically tag the BCE Test Suites with the relevant NER category. We kept the other three categories, *Agreement*, *Register* and *Untranslated*, as-is. The final number of test sentences per LP and category can be found in Table 2.

4.2 Machine Translation

The MT output analyzed in this work was generated using a variety of proprietary MT systems developed by Unbabel. These MT engines are based on Transformer models (Vaswani et al., 2017) and trained using the Marian toolkit (Junczys-Dowmunt et al., 2018). The extent of domain adaptation varies depending on aspects, such as the LP, client, and intended use case (e.g., chat or emails). The generic engines used as the base for domain adaptation are trained on millions of sentences of publicly available parallel data from various domains, for example news, while domain-specific models are fine-tuned on tens to hundreds of thousands of parallel sentences of proprietary content. Models undergo periodic and frequent retrainings[3] to account for domain shift. Not all retrained models enter production right after the training. To decide if a newly trained model should replace the model that is in production during that time, a quality assessment is performed using COMET (Rei et al., 2020) to compare the overall quality of both models. Parallel to this, the available BCE Test Suites are also used for the newly created model and the obtained scores are stored in a database and made accessible and visible to the rest of the company through a BI platform.

For the purpose of this paper, we will showcase two newly introduced improvements in the MT environment. Firstly, we leveraged Factors technology (Dinu et al., 2019; Coelho, 2021) to improve glossary (i.e., clients' terminology) handling of our models. Furthermore, a change in our infrastructure allowed us to easily use in the training environment new entity handling techniques such as better NER models, more refined NER detection and localization strategies that we were already using in production. In Section 5.1 we will show how the BCE Test Suites proposed here are key for validating the improvements obtained by the introduction of the new features mentioned above.

4.3 Quality Estimation

We measure the Business Critical Errors recall using the BCE Test Suites of two separate QE systems developed by or in partnership with Unbabel. The first is a system fine-tuned on Unbabel's proprietary MQM annotation data, and is designed to predict pure MQM scores with high precision. It is trained with a multitask objective and produces token-level OK/BAD tags in addition to sentence scores. The fine-tuning data consists of several million examples, distributed across several dozen LPs, all with English source. The model is based on the OPENKIWI (Kepler et al., 2019) framework and is fine-tuned on the multilingual pre-trained language model XLM-RoBERTa (Conneau et al., 2020).

The second system was developed for the 2022 WMT Shared Task in Quality Estimation (Zerva et al., 2022). Specifically, it is the MQM model listed in Table 3 of Rei et al. (2022) labeled *Word-level + Sentence-level + LP prefix + APEQuest & QT21 + tuned class-weights*. It is a multilingual system based on InfoXLM (Chi et al., 2021), and it is trained with the multitask objective. The system and the COMETKIWI framework with which it is built are described in more detail in Rei et al. (2022).

[3]Using Apache Airflow (`https://airflow.apache.org/`) as the workflow manager.

	LP							
Category	**en-de**	**en-es**	**en-es-latam**	**en-fr**	**en-it**	**en-pt-br**	**en-ru**	**en-zh-cn**
City	69	333	117	285	170	138	179	56
Country	118	332	264	209	136	125	317	87
Currency	229	426	135	141	123	66	156	-
Date	87	161	538	128	208	50	255	-
PRS/ORG	129	371	653	790	399	215	429	72
Agreement	173	85	275	112	144	-	220	N/A
Register	-	-	-	-	-	-	1550	-
Untranslated	-	112	120	84	-	108	-	-

Table 2: Total number of test sentences for the subset of BCE Test Suites. For en-zh-cn *Agreement* Test Suites are not available as this error type does not apply to this language.

5 Results

Our goal was to evaluate the performance of the MT outputs after the new implementations mentioned in Section 4.2 and the BCE recall of the new QE systems mentioned in Section 4.3. The results will be outlined in the Sections below.

5.1 Machine Translation Results

Figures 1 and 2 showcase examples of how the BCE Test Suites can be used to monitor quality across MT models, but also how they can be used for regression purposes of single models.

Figure 1 shows, over time, the average of scores obtained for our domain-adapted models when evaluated on each of the available BCE Test Suites (due to the high cardinality of models across different LPs, an average was preferred over multiple individual charts). In this figure, the highlighted areas (i.e., Factors and Improved entity handling) represent moments when the evaluation in the BCE Test Suites revealed the significant impact on models' performance of the two improvements introduced:

- Factors were leveraged to improve glossary handling, as explained in Section 4.2. This change was gradually launched to all LPs in a span of four months. The ascending trend of the *Agreement* score during this period allows us to observe the positive impact this change had in this type of entity handling.

- As explained in Section 4.2, we started to use several new features for different types of entity handling and detection (e.g., better NER models and localization strategies). This change was done in the 8th month covered in the chart of Figure 1, and the boost in scores during this month for *City*, *Country* and *PRS/ORG*, indicates how much the engines improved in handling these types of entities.

Without the possibility of using the BCE Test Suites for MT evaluation, the impact of both these features could have been obscured when relying solely on automatic metrics that evaluate overall quality, hence the importance of having this type of test set as an extra source of information.

Besides highlighting the impact of new added features, the BCE Tests Suites also allow us to have a historical view of the performance of models in key aspects of the business. We can easily infer if our models changed slightly or decreased their performance on the handling of a certain entity over time, and take actions to counter these behaviors accordingly.

Since BCE Test Suites scores are registered for each retraining iteration, it is also possible to zoom-in into each of the models to obtain a figure like Figure 2 where the scores on the BCE Test Suites for consecutive versions of a model are represented. The shapes around each model version number represent if that model version was deployed to production (green square) or not deployed (red circle).

From Figure 2, it is possible to conclude the following insights:

- Firstly, we can verify how, historically, this model has performed regarding what is evaluated in each BCE Test Suites. We can conclude how we improved for *City*, *Country*, *Date* and *Untranslated*, remained stable for *Currency*, and slightly decreased for *Agreement* and *PRS/ORG*. During a model's life cycle, we can see how scores fluctuate (and not always positively). Since these models live in dynamic environments, small features from other systems can have a significant impact on the quality of the model (e.g., a change in a

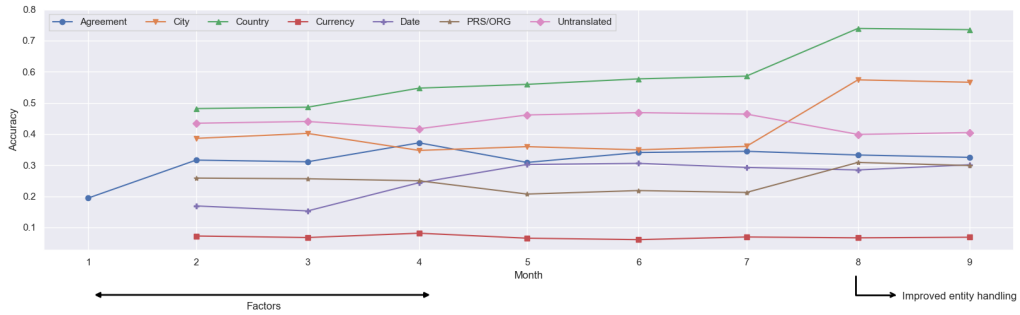

Figure 1: Average score on BCE Test Suites across domain-adapted models.

Figure 2: BCE Test Suites scores for a single customer-adapted model across its several versions. (A green square around the model version indicates that the model version was deployed to production while a red circle indicates it was not.)

tokenization rule) that we would not be aware of without the information provided from the BCE Test Suites. That is why having this view is essential as these insights can be followed by actions and improvements on the systems;

- Secondly, it is noticeable how scores for *City*, *Country* and *Date* Test Suites increased considerably from Version 1 to 2, despite Version 2 not being deployed to production. At Unbabel, we are not actively using these Test Suites for the deployment decision, but instead rely solely on automatic metrics, like COMET (Rei et al., 2020). However, examples such as this show the importance of factoring these scores into the deployment decision. For some clients, it might be more important to avoid mistranslating certain entities, therefore benefiting from having a model in production that performs better in a specific BCE Test Suite and does not compromise the overall quality. For example, industries like travel might require high accuracy on cities, countries and dates, whereas an industry like finance might prioritize accuracy on numbers and currencies;

- Finally, from Version 4 onward all models were deployed to production, which means that the model improved or did not degrade in the automatic metric scores. The same can be said for the BCE Test Suites scores. The desired behavior is that these scores plus automatic metrics can be used together to perform a more realistic and trustworthy deployment decision. This could increase the confidence that the new model is equally good or better both in terms of automatic metrics (measuring average quality) but also in BCE Test Suites (measuring important business metrics).

All these insights are only possible when using different types of test sets that can measure different features and details of the translations. These allow us to monitor and track how quality changes over time, but also how new features can have an impact on the engines' performance.

5.2 Quality Estimation Results

Table 3 shows BCE recall results for the two QE systems described in Section 4.3. Overall, the MQM-QE model consistently outperforms the WMT model. This is not surprising since the MQM-QE model was fine-tuned with millions of

	LP								
Error Category	**en-de**	**en-es**	**en-es-latam**	**en-fr**	**en-it**	**en-pt-br**	**en-ru**	**en-zh-cn**	**Cat. avg.**
				MQM-QE					
Agreement	0.932	0.991	-	0.976	-	-	0.992	-	0.972
City	0.522	0.898	0.592	0.627	0.787	0.487	0.774	0.621	0.663
Country	0.575	0.790	0.845	0.846	0.802	0.564	0.799	0.718	0.742
Currency	-	0.972	0.762	0.976	0.303	0.276	0.788	-	0.679
Date	-	0.852	0.669	0.933	0.813	0.833	0.912	-	0.835
PRS/ORG	-	0.531	0.468	0.589	-	0.279	0.588	-	0.491
Register	-	-	-	-	-	-	0.984	-	0.984
Untranslated	-	0.814	0.754	0.881	-	0.759	-	-	0.802
LP avg.	0.676	0.835	0.682	0.833	0.676	0.533	0.834	0.669	0.717
				WMT-word-level-QE					
Agreement	0.749	0.914	-	0.888	-	-	0.745	-	0.824
City	0.356	0.879	0.583	0.455	0.711	0.470	0.429	0.690	0.572
Country	0.500	0.731	0.744	0.803	0.648	0.594	0.473	0.732	0.653
Currency	-	0.628	0.752	0.554	0.382	0.652	0.311	-	0.547
Date	-	0.260	0.279	0.600	0.813	0.611	0.391	-	0.492
PRS/ORG	-	0.490	0.468	0.600	-	0.552	0.350	-	0.492
Register	-	-	-	-	-	-	0.077	-	0.077
Untranslated	-	0.559	0.435	0.607	-	0.435	-	-	0.509
LP avg.	0.535	0.638	0.543	0.644	0.638	0.553	0.396	0.711	0.582

Table 3: BCE recall results for MQM-QE and WMT word-level QE model.

examples of Unbabel-MQM data, which matches the domain of the Test Suites. The WMT model, on the other hand, was fine-tuned with publicly-available generic data, out-of-domain for the Test Suites. Given this, the WMT model does remarkably well, especially considering that the MQM data for fine-tuning only included three LPs: "English–German", "English–Russian", and "Simplified Chinese–English".

The BCE recall analysis is also useful for highlighting specific areas of strength and weakness for the MQM-QE model. One of its main strengths is flagging instances of the incorrect register or tone.[4] *Register* is an important component of the MQM typology, especially in the Customer Service domain. The MQM-QE model scores nearly perfectly in this category, while the WMT model barely detects any errors. This suggests that specializing fine-tuning or training data to the business use case gives improvement over using more generic systems out-of-the-box, and that there is value in leveraging domain- or use case-specific expertise.

The BCE Test Suites are also able to indicate that the MQM-QE model could be improved in detecting certain named-entity errors: *Currency* for "English–Italian" and "English–Brazilian Portuguese", *City* for "English–German" and "English–Brazilian Portuguese", and *Products-Organizations* for "English–Spanish-Latam" and "English–Brazilian Portuguese". This suggests that more investigation into the fine-tuning data is required, as it is possible that we are lacking in data for these categories, or that the annotations of these errors are inconsistent. This kind of analysis, however, is only made possible in a scalable way by the BCE Test Suites. The evaluation is *actionable* and opens up avenues for model improvement whose necessity was not obvious before, such as data cleaning and data augmentation.

6 Conclusions and Future Work

In this work, we present a methodology to build Test Suites that are tailored to address Business Critical Errors (Stewart et al., 2022) and how they could potentially harm customers in a business setting.

We demonstrated that it is possible to use a dataset of translation errors annotated by following the MQM framework (Lommel et al., 2014) of 'real-life' machine translation errors to build comprehensive Test Suites for several LPs in order to evaluate the performance of both MT and QE systems.

As shown in Section 5.1, relying on the BCE Test Suites scores alongside the automatic metrics to decide whether a model should be deployed to

[4]Due to time restrictions, we currently only have a *Register* Test Suite for "English–Russian"; adding more LPs in this category is high priority for future work.

production or not brings great value to the robustness of the model, hence results about BCE Test Suites accuracy will be added to the automatic deployment criteria.

BCE Test Suites are also a valuable part of the QE evaluation pipeline, highlighting errors that are important in a business setting.

For future work, we would like to extend the information in the Test Suites to include all errors in the sentence, so we can measure precision-based metrics as well. Table 2 shows the number of BCE Test Suites available per LP and category and it can be seen that for some LPs there is the need to create full sets of test sentences, which is already a work in progress.

Finally, we aim to extend the BCE Test Suites to more LPs and language varieties that were not previously addressed, namely "English–Japanese", "English–Korean", "English–Portuguese" and "English–Traditional Chinese". The second goal is to have more Test Suites dedicated to more BCE categories, such as *Locale Conventions* issues.

Acknowledgements

The authors want to thank Beatriz Silva, Tânia Vaz, Natalia Sugrobova, Sandra Rosa, Teresa Marmeleira and Katherine Zhang for their crucial help in the BCE Test Suites' production. Besides this, the authors also want to thank Craig Stewart, whose expertise about metrics and evaluation was paramount for the development of the BCE Test Suites, and Amin Farajian, for his great support and knowledge about Machine Translation. This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 and by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

References

Alves, Duarte, Ricardo Rei, Ana C Farinha, José G. C. De Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Avramidis, Eleftherios, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA, March. Association for Machine Translation in the Americas.

Avramidis, Eleftherios, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2019. Fine-grained evaluation of quality estimation for machine translation based on a linguistically-motivated test suite. *CoRR*, abs/1910.07468.

Banerjee, Satanjeev 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.

Chi, Zewen, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June. Association for Computational Linguistics.

Coelho, Pedro Dias. 2021. Factored Models for Neural Machine Translation. Master's thesis, Instituto Superior Técnico.

Conneau, Alexis, 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, July. Association for Computational Linguistics.

Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July. Association for Computational Linguistics.

Fonseca, Erick, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

Guillou, Liane, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium,

Brussels, October. Association for Computational Linguistics.

Isabelle, Pierre, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark, September. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Kepler, Fabio, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*.

Lommel, Arle, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463, 12.

Macketanz, Vivien, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France, June. European Language Resources Association.

Matthews, Brian W. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Menezes, Miguel, Vera Cabarrao, Pedro Mota, Helena Moniz, and Alon Lavie. 2022. A case study on the importance of named entities in a machine translation pipeline for customer support content. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 211–219, Ghent, Belgium, June. European Association for Machine Translation.

Mota, Pedro, Vera Cabarrao, and Eduardo Farah. 2022. Fast-paced improvements to named entity handling for neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 141–149, Ghent, Belgium, June. European Association for Machine Translation.

Papakipos, Zoe and Joanna Bitton. 2022. Augly: Data augmentations for robustness.

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

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Specia, Lucia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain, May 14–15. European Association for Machine Translation.

Stewart, Craig A, Madalena Gonçalves, Marianna Buchicchio, and Alon Lavie. 2022. Business critical errors: A framework for adaptive quality feedback. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 231–256, Orlando, USA, September. Association for Machine Translation in the Americas.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zerva, Chrysoula, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.