Towards User-Reliable Machine Translation Evaluation

Sofía García González[1,2]

[1]University of the Basque Country (UPV/EHU), Manuel Lardizabal pasealekua, 1, 20018 Donostia-San Sebastian, Gipuzkoa (Spain)
[2]**imaxin**|software, Rúa Salgueiriños de Abaixo, 11, Santiago de Compostela, 15706, Galicia (Spain)

Abstract

The advent of Neural Machine Translation has ushered in a new era of progress in the field of Natural Language Processing. However, Neural Machine Translation is not without its limitations. It is prone to a number of translation errors, including omissions, hallucinations, and other issues that arise from the machine translation process itself. These errors can be problematic in production environments. This thesis will evaluate different methodologies for automatic Machine Translation evaluation and error detection for low resource languages in an industrial context.

Keywords

Machine translation, Automatic Evaluation, Low Resource Languages

1. Introduction and Motivation

This research project is a part-time industrial thesis between the University of the Basque Country (UPV/EHU) and **imaxin**|software, a software company from Galicia specialized in Machine Translation (MT), specially for Spanish, Galician and other LRL. The main topic of this thesis is MT evaluation.

MT can be defined as any use of computer systems to transform a computerised text written in a source language into a different computerised text written in a target language, thereby generating what is known as a raw translation [1]. This field has evolved tremendously since the advent of transformer models in 2017, as well as other text generation tasks [2]. Neural Machine Translation (NMT) generates more fluent, more natural and more accurate translations than older methods such as Rule-Based Machine Translation (RBMT) or Statistical Machine Translation (SMT), even between distant languages [3]. Additionally, the advent of Large Language Models (LLM) has precipitated a paradigm shift in NMT, giving rise to novel scenarios that evolve at a rapid pace [4]. However, the errors made by these systems are challenging for users to detect, particularly because the most significant errors are not translation errors, but errors in the models themselves. This renders them unreliable for both users and companies that wish to implement them. This is why MT evaluation methods are of such importance [5].

The MT evaluation can be divided into two categories: human evaluation and automatic evaluation, which has traditionally involved metrics that compare a reference text with the translation hypothesis, resulting in a corpus-based metric. However, until the advent of recent studies, there has been no automatic evaluation method capable of detecting and pointing out translation errors due to the intrinsic difficulty of this task [5].

The novel approaches to MT and MT evaluation are predominantly framed within the context of High Resource Languages (HRL) such as English. The substantial quantity of data, often annotated, required to train these types of evaluation systems, or the absence of a basic neural or LLM model, render these methods ineffective or even detrimental in LRL, such as Galician. This is the rationale behind this thesis, which seeks to address the following research question: **"Which is the most effective method for detecting errors and evaluating machine translation in an industrial context for low-resource languages?"**.

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.
sofia.garcia@imaxin.com (S. G. González)

2. Related Work

In recent years, Natural Language Processing (NLP) has improved and evolved significantly due to the new architectures and higher computational power [2]. NMT has greatly improved the performance of RBMT and SMT, particularly after the development of the Transformer architecture [3]. And, more recently, even LLMs as GPT-3[1] or GPT-4[2] have also shown promising results in this field, both for generic translation [6] and domain adaptation [7].

2.1. Machine Translation Errors

Nowadays, machine translation has become more sophisticated, more fluent and, the advent or multilingual models, have facilitated the advancement and deployment of MT between distant and LRL languages. Nevertheless, despite the aforementioned advances, machine translation remains imperfect, particularly due to the fact that the new type of errors made by these systems are often not identified by the user, despite their potential for causing significant issues. These errors can be categorised into two distinct types. The first category encompasses errors intrinsic to the **system** itself, such as **hallucinations** and **omissions**, whereas the second category comprises **translation errors**, including those pertaining to **grammar**, **syntax** and **style** [8].

Hallucinations can occur during translation when the model produces a sentence that is grammatically correct but does not accurately convey the meaning of the source sentence. These can be either partial, where only a portion of the source sentence is inaccurately translated, or total, where the entire translated sentence is incorrect in content [8]. On the other hand, omissions can occur if the model forgets part of the source sentence, resulting in the loss of relevant information from the original sentence [9]. These are the most significant errors, as they can result in the loss of information or a change of meaning in the machine translation. Conversely, syntactic and morphological errors, as well as the incorrect translation of entities, are inherent to machine translation and may also occur in this type of models.

The causes of errors intrinsic to neural systems, such as hallucinations or omissions, remain unsolved. Although these phenomena do not occur generically in machine translation, they can give rise to misunderstandings and problems in its production. With regard to the errors inherent in machine translation, it is possible that many may be due to the training corpus itself [8].

2.2. Machine Translation Evaluation

Nowadays MT evaluation can be divided into two main categories: **human evaluation** and **automatic evaluation**. Human evaluation, while more dependable, is more expensive and time-consuming. It assesses the **adequacy** of the translation, ensuring that the meaning of the original sentence is preserved in the translated one, as well as the **fluency**, ensuring that the translation is grammatically and syntactically correct in the target language. Within this domain, there exist a variety of evaluation frameworks. These frameworks serve as a guide for evaluators, providing a consistent approach to the MT evaluation. Two prominent examples are the Multidimensional Quality Metrics (MQM) and the Direct Assesments (DA). These type of annotations have been used to train different evaluation models [10].

2.2.1. Automatic Evaluation: Metrics

On the other hand, automatic evaluation is less costly and time-consuming, but less reliable than human evaluation. Traditionally, the MT system quality has been evaluated with **lexical based** metrics. These metrics provide a score at the sentence or document level by comparing the MT output with a reference text previously reviewed by linguists or native speakers, named gold standard. This comparison between the gold standard and the MT output can be done at three levels: at token level, using metrics that

[1] https://chatgpt.com/
[2] https://openai.com/index/gpt-4/

compare both documents token by token, such as Bilingual Evaluation Understudy (BLEU) [11], which has been the reference metric until nowadays; at character level, such as CHaRacter-level F-score (chrF) [12]; and, finally, by measuring the number of changes required to transform the MT output into the reference text such as: insertions, deletions, substitutions or rearrangements. This metric is known as Translation Edit Rate (TER) [13]. Such metrics have been the subject of significant criticism, with two key factors being identified as the primary sources of it. Firstly, they are reliant on a reference text with which to compare the MT output, which is not always feasible in all contexts. Secondly, natural language is highly versatile, thus they may evaluate as erroneous translations that are in fact correct, even if they differ from the gold standard [14].

More recently, metrics based on **embeddings** have emerged. These metrics extend beyond a mere lexical comparison of translations and references, instead enabling a semantic comparison at the word or sentence embedding level. An example of this type of metric is the Crosslingual Optimized Metric for Evaluation of Translation (COMET), a multilingual MT evaluation framework that is able to evaluate MT with different reference-dependent and reference-free models [10].

Another method for automatic MT evaluation is **Qualitative Estimation (QE)**. It is the task of estimating the quality of MT in real time without the need for reference translations [15]. Some examples of QE metrics are TransQuest[3] [16, 17], a bilingual and multilingual framework that uses XLM-R embeddings of source and hypothesis sentences to predict the QE score; wmt-comet-qe-da and wmt-cometkiwida models[4] [18]. Although there is an advantage in not requiring a reference text for evaluation, QE is often limited to HRL and general domains, as it requires an annotated training corpus. This MT evaluation can be conducted at the word, sentence, or document level [15].

2.2.2. Automatic Evaluation: Error Detection

In recent times, a number of studies have been conducted with a view to examining the phenomenon of error detection in greater depth. For instance, LLMs have shown a high capacity to evaluate machine translation through error analysis, pointing out errors and even explaining them and suggesting improvements to the translation. However, they still fail to correlate error analysis with translation quality metrics [19]. Despite the good quality shown by big LLMs as GPT3 or GPT4, when attempts have been made to replicate this task in open-source LLMs such as LLAMA,[5] the results have deteriorated [20].

On the other hand, Unbabel,[6] the COMET development company, has released a new model (XCOMET[7]) which is capable of detecting linguistic errors and hallucinations and classifying them according to their severity: minor, serious or critical [20]. Nevertheless, both the LLMs and XCOMET have only been tested in HRL, such as English, Chinese or Russian. LRL have less presence in large language models or less training data for training this types of systems.

In this context, Dale et al. [9] have produced manually annotated datasets to detect hallucinations and omissions. These datasets encompass 18 language pairs, including HRL pairs, LRL pairs, and one zero-shot pair. This work aims to advance the field of hallucination and omission detection and is the first one to produce a human annotated dataset to this purpose.

Finally, another significant contribution to this field is that of Don-Yehiya et al. [21]. In this work they have investigated the possibility of predicting the quality of a translated sentence based on the semantic and linguistic characteristics of the source sentence. To this end, they have developed a model capable of determining the intrinsic difficulty of a source sentence. In this work, they have been able to conclude that the characteristics of a sentence determine the difficulty or ease of translation for different machine translation models. Consequently, it is now possible to predict whether a sentence is going to be mistranslated before it occurs. This can also be of great assistance to both linguists and

[3] https://huggingface.co/TransQuest/monotransquest-da-en_de-wiki
[4] https://github.com/Unbabel/COMET/blob/master/MODELS.md
[5] https://llama.meta.com/
[6] https://unbabel.com/
[7] https://unbabel.com/xcomet-translation-quality-analysis/

users. But more investigation is needed.

In conclusion, research on error detection in machine translation or generative tasks is still in its infancy and is only being developed for HRL. For other contexts such as LRL and very specific domains, there is very little training corpus or presence of these languages in the base models. Moreover, both the language models and the methods used by neural models such as COMET require a high computational cost, which makes them inefficient methods to put into production in small companies.

3. Research Questions and Hypotheses

3.1. Evaluation Metrics

RQ1: What is the starting point? In order to assess the value of the information that some of the metrics mentioned in section 2.2.1 can provide, the initial step of this thesis will be to establish, for the first time, the state of the art in machine translation for Spanish–Galician and English–Galician pairs. The first part of this research question has been published at PROPOR 2024.[8] In this first paper, we have conducted an exhaustive evaluation of all extant models for these language pairs, encompassing both NMT and RBMT systems, across several test datasets pertaining to the general, legal, and health domains. The metrics employed were BLEU, chrF, TER, and COMET, in addition to an error analysis [22]. The remaining translation direction will be published in due course.

The results obtained in this initial study indicate a correlation between the lexical-based metrics. This implies that they exhibit homogeneous results, with minimal variation between them, and consistently follow the same patterns across models. In contrast, COMET yields considerably higher results than the lexical-based metrics, with smaller differences between models. From these initial findings, it can be concluded that the metrics are useful for identifying models that exhibit poor performance, indicating the need for further revision and retraining. Conversely, models with high metrics scores demonstrate satisfactory performance, although the metrics themselves provide limited insight into the nuances of translation quality.

RQ2: If each sentence is individually evaluated, can we identify a correlation between the metrics and the errors? Once the corpus-level evaluations have been completed, another question that must be addressed is whether the sentence-by-sentence evaluation aligns with the error analysis. This entails determining whether and how the metrics penalise the same types of errors. To answer this question, the same metrics mentioned in RQ1 will be employed to evaluate the generic corpus on a sentence-by-sentence basis. The primary hypothesis is that, contingent on the typology (lexical-based, edit distance-based or embedding-based), each metric will impose a distinct penalty on the errors identified in the sentences. Consequently, each metric will yield disparate insights.

RQ3: Can errors in machine translation be identified prior to their occurrence? As evidenced by the Don-Yehiya et al. [21] work and the analysis made in García and Rigau [22], which revealed that each translation model consistently produces similar errors, our hypothesis is that it is feasible to develop a model that can predict the likelihood of certain linguistic phenomena being challenging to detect in the target language, contingent on the translation model employed.

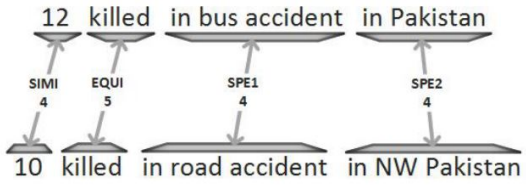
3.2. Error Detection in the Machine Translation

RQ4: What is the most appropriate methodology for the generation of annotated data in a LRL? One of the significant limitations of a low-resource language such as Galician for the training of systems capable of detecting errors is the lack of an annotated corpus. Works such as XCOMET [20], PreQuel [21] or the LLM themselves have required an annotated corpus with errors for their training. In the case of Galician, it would be necessary to generate this kind of data from scratch. Consequently, we will adopt the methodologies proposed in Guerreiro et al. [20] to tag data in accordance with the characteristics of the MQM corpora, or the methodology proposed by Dale et al. [9] to generate annotated corpora with hallucinations and omissions.

[8] https://propor2024.citius.gal/

RQ5: Is it possible to detect MT errors through interpretable Semantic Textual Similarity (iSTS)? Semantic textual similarity is defined as the measure of semantic equivalence between two blocks of text [23]. In the WMT12, Castillo and Estrella [24] proposed, for the first time, the use of semantic textual similarity as machine translation evaluation. They proposed the use of WordNet to calculate how much equivalent is a hypothesis sentence to a reference. They achieved, at that year, promising results at system level. Nowadays, metrics as COMET or BLEURT [25] use sentence and word embeddings to compare the semantic similarity at word or sentence level.

In contrast, to the best of our knowledge, there has never been a proposal for the use of interpretable semantic textual similarity in MT evaluation and error detection. iSTS can be defined as giving meaning to semantic similarity between short texts [26]. See the example in Figure 1.



(a) The following two sentences, taken from Lopez-Gazpio et al. [27] provide an explanation of the interpretability layer between two similar sentences in the English language.

The two sentences are very similar. Note that 'in bus accident' is a bit more specific than 'in road accident' in this context. Note also that '12' and '10' are very similar in this context. Note also that 'in Pakistan' is a bit more general than 'in NW Pakistan' in this context.

(b) Explanation given by the model in Lopez-Gazpio et al. [27] about the differences between the two sentences in subfigure 1a.

Figure 1: Figures taken from Lopez-Gazpio et al. [27] article that explain the interpretable Semantic Textual Similarity model approach.

In Subfigure 1a, *12 killed in bus accident in Pakistan* and *10 killed in road accident in NW Pakistan* differ from each other in two minor respects: the number of people killed and the type of accident. Although the meaning conveyed by the two sentences is similar, it differs between them. It is precisely these aspects in which the two sentences differ that are highlighted in the text of the Subfigure 1b [27]. This thesis therefore proposes the extension of iSTS to a multilingual and crosslingual context as an evaluation approach for machine translation. In this context, the interpretability layer will function as an error analysis that is able to explain the MT errors to the user. The aim is therefore to develop a metric capable of identifying discrepancies between languages without the need for a reference test. The hypothesis is that, with the current development of neural networks and large language models, it will be possible to expand iSTS to a multilingual context. The objective is to optimise this technique to achieve a metric similar to XCOMET, but capable of explaining errors and not just classifying them.

RQ6: Can open-source LLM identify errors in LRL contexts? As previously stated in section 2.2.2, despite the remarkable capacity of large language models, such as GPT3 and GPT4, to conduct error analysis on MT evaluation [19], Guerreiro et al. [20] have also indicated that open-source LLM exhibit inferior performance and fail to attain the quality of private LLM.

Currently, there are two LLM available for Galician: Carballo-bloom-1.3B[9] and Carballo-cerebras-1.3B.[10] In spite of that, none of them are fine-tuned to MT, MT evaluation or other specific tasks. We will conduct in this thesis a fine tuning of Carballo-bloom-1.3B, as it is a model based on FLOR-1.3B[11] that has been also based on the multilingual LLM Bloom-1.7B.[12] Consequently, it is anticipated that the ability to recognise languages other than Galician, such as English or Spanish, will be demonstrated. The advent of enhanced techniques for enhancing these models, even in the context of low-resource languages, suggests that even modest LLM may yield satisfactory performance.

[9] https://huggingface.co/proxectonos/Carballo-bloom-1.3B
[10] https://huggingface.co/proxectonos/Carballo-cerebras-1.3B
[11] https://huggingface.co/projecte-aina/FLOR-6.3B-Instructed
[12] https://huggingface.co/bigscience/bloom-1b7

3.3. Error Correction

The objective of this study is not only to identify errors in machine translation but also to ascertain the feasibility of correcting them. The following section presents two hypotheses that have been formulated in this regard.

RQ7: Can we use multilingual models to translate from one bad MT to a good MT? One of the research questions to be addressed in this thesis is whether it is possible to use multilingual models to translate from one language into the same language in order to improve the MT made by another model. That is, to provide the model with an MT with errors and have it translate it into the same language, in order to rectify the existing errors. The hypothesis is that a multilingual model with sufficient knowledge of the language will be able to correct a machine translation generated by another model.

In order to achieve this objective, we will be utilising the multilingual models M2M100[13] [28] and NLLB200[14] [29], which include Galician among their languages and have been identified as the most effective in García and Rigau [22].

RQ8: Is it possible to improve translation models based on error detection? The hypothesis put forth for this research question is that every model will make certain types of errors. Therefore, detecting these errors will not only assist the user, but also serve as a basis for retraining in order to improve the model.

4. Timeline

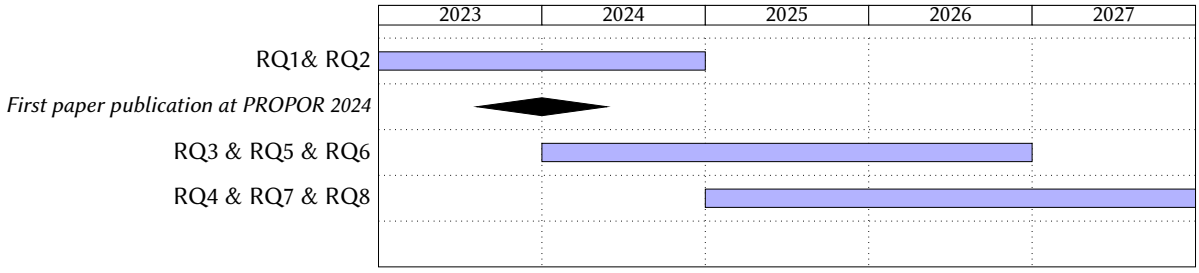


Figure 2: Thesis Timeline

The Figure 2 illustrates the planning of the thesis. The initial part will encompass an examination of conventional MT evaluation metrics and will mark the significant shift occurring in the realm of MT and its evaluation as a result of the advent of LLM. A preliminary paper on this aspect of the research has already been published as it has been mentioned in section 3 for RQ1. By the end of 2024, the initial section of the thesis is anticipated to be completed, addressing RQ1 and RQ2.

On the other hand, as previously discussed in section 3, the publication of the new LLMs for Galician will be accompanied by experiments aimed at retraining them for MT and MT evaluation. To this end, it will be necessary to create an annotated corpus with errors and their typology, which will address research RQ4 and RQ6. Similarly, the creation of these corpora will permit the training of systems based on embeddings to detect errors, thereby providing an answer to RQ5. Furthermore, the detection of errors prior to translation will be facilitated, thus providing an answer to RQ3. The commencement of this part of the thesis is scheduled to take place between mid-2024 and mid-2026. Finally, the final year will be dedicated to enhancing the systems themselves by identifying errors (RQ8) or improving the translation process (RQ7).

[13] https://huggingface.co/facebook/m2m100_418M
[14] https://huggingface.co/facebook/nllb-200-distilled-600M

5. Conclusions

In conclusion, the aforementioned experiments in Section 3 will be employed in order to ascertain the optimal methodology for MT evaluation for a LRL in an industrial context. Given the rapidly evolving environment in which we have been living in recent years, the new technologies that are being generated tend to be focused on HRL and require a very high computational cost for SMEs. The objective of this thesis is to conduct a comparative study of the existing technologies and to assess their potential for implementation in a real-world context.

Acknowledgments

We would like to express our gratitude to the Nós project members for their assistance and guidance during the development of the methodological part of the project. Additionally, computational resources for this research were provided by UPV/EHU and **imaxin**|software. Finally, we acknowledge the funding received from the following projects:

(i) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe.

(ii) DeepR3 (TED2021-130295B-C31) and European Union NextGeneration EU/PRTR.

(iii) ILENIA (2022/TL22/00215335) the EU-funded NextGenera- tionEU Recovery, Transformation and Resilience Plan.

References

[1] M. L.Forcada, Building machine translation systems for minor languages: Challenges and effects, Revista de llengua i dret // Journal of language and law (2020).

[2] D. W. Otter, J. R. Medina, J. K. Kalita, A survey of the usages of deep learning for natural language processing, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 604–624. doi:10.1109/TNNLS.2020.2979670.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[4] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, L. Li, Multilingual machine translation with large language models: Empirical results and analysis (2023). URL: http://arxiv.org/abs/2304.04675.

[5] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, H. Lim, A survey on evaluation metrics for machine translation, Mathematics 11 (2023). doi:10.3390/math11041006.

[6] B. Zhang, B. Haddow, A. Birch, Prompting large language model for machine translation: A case study, in: International Conference on Machine Learning, PMLR, 2023, pp. 41092–41110.

[7] Y. Moslem, R. Haque, J. D. Kelleher, A. Way, Adaptive machine translation with large language models, in: M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, H. Moniz (Eds.), Proceedings of the 24th Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Tampere, Finland, 2023, pp. 227–237. URL: https://aclanthology.org/2023.eamt-1.22.

[8] W. Xu, S. Agrawal, E. Briakou, M. J. Martindale, M. Carpuat, Understanding and detecting hallucinations in neural machine translation via model introspection, Transactions of the Association for Computational Linguistics 11 (2023) 546–564. URL: https://aclanthology.org/2023.tacl-1.32. doi:10.1162/tacl_a_00563.

[9] D. Dale, E. Voita, J. Lam, P. Hansanti, C. Ropers, E. Kalbassi, C. Gao, L. Barrault, M. Costa-jussà, Halomi: A manually annotated benchmark for multilingual hallucination and omission detection

in machine translation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 638–653.

[10] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213. doi:10.18653/v1/2020.emnlp-main.213.

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[12] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/v1/W15-3049.

[13] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, 2006, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25.

[14] M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, A. F. T. Martins, Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust, in: P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, M. Zampieri (Eds.), Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 46–68. URL: https://aclanthology.org/2022.wmt-1.2.

[15] H. Zhao, Y. Liu, S. Tao, W. Meng, Y. Chen, X. Geng, C. Su, M. Zhang, H. Yang, From handcrafted features to llms: A brief survey for machine translation quality estimation, arXiv preprint arXiv:2403.14118 (2024).

[16] T. Ranasinghe, C. Orasan, R. Mitkov, Transquest: Translation quality estimation with cross-lingual transformers, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020.

[17] T. Ranasinghe, C. Orasan, R. Mitkov, Transquest at wmt2020: Sentence-level direct assessment, in: Proceedings of the Fifth Conference on Machine Translation, 2020.

[18] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, A. F. T. Martins, CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task, in: P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, M. Zampieri (Eds.), Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 634–645. URL: https://aclanthology.org/2022.wmt-1.60.

[19] Q. Lu, B. Qiu, L. Ding, L. Xie, D. Tao, Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt (2023).

[20] N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, A. F. T. Martins, xcomet: Transparent machine translation evaluation through fine-grained error detection, 2023. arXiv:2310.10482.

[21] S. Don-Yehiya, L. Choshen, O. Abend, Prequel: Quality estimation of machine translation outputs

in advance (2022). URL: http://arxiv.org/abs/2205.09178.

[22] S. García, G. Rigau, Study of the state of the art Galician machine translation: English-Galician and Spanish-Galician models, in: P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, R. Amaro (Eds.), Proceedings of the 16th International Conference on Computational Processing of Portuguese, Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, 2024, pp. 411–421. URL: https://aclanthology.org/2024.propor-1.42.

[23] D. Chandrasekaran, V. Mago, Evolution of semantic similarity – a survey (2020). URL: http://arxiv.org/abs/2004.13820http://dx.doi.org/10.1145/3440755. doi:10.1145/3440755.

[24] J. Castillo, P. Estrella, Semantic textual similarity for MT evaluation, in: C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, L. Specia (Eds.), Proceedings of the Seventh Workshop on Statistical Machine Translation, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 52–58. URL: https://aclanthology.org/W12-3103.

[25] T. Sellam, D. Das, A. P. Parikh, Bleurt: Learning robust metrics for text generation, in: Proceedings of ACL, 2020.

[26] A. A. Abafogi, Survey on interpretable semantic textual similarity and its applications, International Journal of Innovative Technology and Exploring Engineering 10 (2021) 14–18. URL: https://www.ijitee.org/portfolio-item/B82941210220/. doi:10.35940/ijitee.B8294.0110321.

[27] I. Lopez-Gazpio, M. Maritxalar, A. Gonzalez-Agirre, G. Rigau, L. Uria, E. Agirre, Interpretable semantic textual similarity: Finding and explaining differences between sentences (2016). URL: http://arxiv.org/abs/1612.04868http://dx.doi.org/10.1016/j.knosys.2016.12.013. doi:10.1016/j.knosys.2016.12.013.

[28] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al., Beyond english-centric multilingual machine translation, Journal of Machine Learning Research 22 (2021) 1–48.

[29] N. team, M. Costa-jussa, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Gonzalez, P. Hansanti, J. Wang, No language left behind: Scaling human-centered machine translation (2022). doi:10.48550/arXiv.2207.04672.