UTER: Capturing the Human Touch in Evaluating Morphologically Rich and Low-Resource Languages

Samy Ouzerrout
Université d'Orléans, France
samy.ouzerrout@etu.univ-orleans.fr

Abstract

We introduce UTER, a novel automatic translation evaluation metric specifically designed for morphologically complex languages. Unlike traditional TER approaches, UTER incorporates a reordering algorithm and leverages the Sørensen-Dicse similarity measure to better account for morphological variations.

Tested on morphologically rich and low resource languages from the WMT22 dataset, such as Finnish, Estonian, Kazakh, and Xhosa, UTER delivers results that align more closely with human direct assessments (DA) and outperforms benchmark metrics, including chrF and METEOR. Furthermore, its effectiveness has also been demonstrated on languages with complex writing systems, such as Chinese and Japanese, showcasing its versatility and robustness.

1 Introduction

With the rise of machine translation systems, evaluating their quality has become a key challenge in numerous fields, particularly for under-represented and morphologically complex languages.

Assessing the quality of machine translations is of critical importance, especially in a context where translation systems are increasingly being adopted for languages that present specific challenges, such as complex morphology.

However, traditional evaluation metrics, such as BLEU (Papineni et al., 2002), rely on simple n-gram matching with reference sentences, making them inadequate for capturing the morphological or syntactic variations characteristic of these languages.

As highlighted by Haddow et al. (2022), these approaches face significant limitations in low-resource language environments, which are often marked by complex morphology. The authors recommend using language models, which are better suited to capture linguistic and semantic nuances.

Nonetheless, language models are generally ill-suited for under-represented languages due to the lack of available data and limitations in contextual representations. This leads to a strong dependency on English and other dominant languages. Moreover, their computational and training time requirements pose significant challenges in resource-constrained settings.

To address these challenges, we propose a new lexical metric, Universal-TER (UTER), designed to enhance the Translation Edit Rate (TER) by incorporating reordering mechanisms and a refined consideration of lexical and morphological variations. Unlike traditional approaches, UTER offers fast, accurate, and resource-independent evaluation.

In this paper, we present UTER's performance by comparing it against human direct assessments (DA) from the WMT dataset, benchmark metrics such as chrF and METEOR, and two alternative metrics. Our results demonstrate that UTER provides a reliable and accurate measure of translation quality, making it particularly valuable for morphologically complex languages.

Before delving into UTER's performance, it is crucial to examine the specific challenges posed by under-represented and morphologically complex languages to evaluation metrics.

An implementation of UTER is publicly available as a Python package on PyPI: https://pypi.org/project/evalnlp/.

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 16–23

2 Challenges and Limitations of Metrics for Morphologically Rich and Under-Represented Languages

Evaluating translations for low-resource languages presents unique challenges, largely due to the complexity of their morphological structures and the scarcity of representative linguistic data. These languages, often spoken within small communities and less influenced by non-native speakers, tend to retain elaborate morphological systems, including variations in suffixes, prefixes, and grammatical inflections (Lupyan and Dale, 2010; Lindenfelser, 2020).

In such contexts, metrics based on lexical matching, such as BLEU and TER, often lack the precision needed to capture the intricate morphological and syntactic nuances of these languages. This makes them particularly unsuitable for languages like Finnish, Kazakh, or Xhosa, where minor morphological variations can result in significant discrepancies in scores, even when such variations are linguistically acceptable.

To address some of BLEU's limitations, the chrF metric (Popović, 2015) was introduced, relying on character-level rather than word-level matches to better capture the morphology of complex languages. While chrF provides an improvement, it remains insufficient for evaluating translation quality with the accuracy of human assessments.

Recent approaches leveraging language models have sought to overcome these issues by capturing contextual and semantic information, thereby going beyond simple word matches. However, due to the limited availability of data for under-represented and morphologically rich languages, these models often lack sufficient examples for proper generalization (Singh et al., 2024).

Moreover, the absence of specific training data limits the accuracy of evaluations, as pretrained models are predominantly optimized for resource-rich languages. This creates a bias that prevents these models from delivering reliable evaluations for under-represented, morphologically rich languages (Lee et al., 2023). Finally, the high computational and training time requirements add an additional hurdle, making these solutions costly and impractical in such settings.

In summary, these limitations highlight the need for an alternative metric tailored to the specificities of morphologically rich and under-represented languages. UTER addresses this need by delivering results that align more closely with human judgments, without relying on advanced language models or large training datasets.

3 Limitations of TER

The *Translation Edit Rate* (TER) evaluates the quality of machine translations by measuring the minimal number of operations (insertion, deletion, substitution, and transposition) required to transform a translation into a reference sentence. Its formula is as follows:

$$\text{TER} = \frac{\text{Total number of operations}}{\text{Length of the reference sentence}}$$

Although TER is a widely used benchmark metric, it has several shortcomings that limit its ability to accurately reflect translation quality:

- **Lack of differentiation between major and minor errors:** TER assigns the same weight to word omission or insertion errors as to transposition or substitution errors, even when some substitutions are semantically equivalent or represent only slight lexical variations.

- **Insensitivity to syntactic variations:** TER ignores the syntactic relationships between words and is limited to exact matching of transpositions and substitutions. This approach makes it unsuitable for grammatically complex structures or languages with flexible syntax.

- **Bias towards short sentences:** TER reports the number of operations relative to the length of the reference sentence, which can introduce bias.

  Dividing by the reference length amplifies penalties for additions in longer translations, while dividing by the translation length accentuates penalties for omissions in shorter translations. While this principle may be justified in ASR, it remains unsuitable for translation evaluation.

  Figure 1 shows how TER varies with normalization choice.

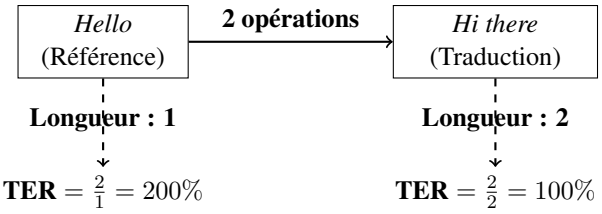

Figure 1: Illustration of the bias introduced by TER's normalization

- **Greedy approach to transposition calculation:**

  This approach successively applies the most beneficial transpositions, those that minimize the number of insertion, deletion, and substitution operations, until no further improvements are possible (Snover et al., 2006).

  While this method allows for an approximate solution, recalculating the edit operations after each transposition results in high computational complexity.

  Furthermore, it prioritizes local improvements without ensuring global optimality, which can lead to suboptimal choices and prevent reaching a more favorable configuration.

  Finally, this strategy ignores the interdependencies between transpositions, where a different combination could further reduce the overall score.

- **Distance of transpositions not taken into account:** Although uniformly penalizing transpositions, regardless of the distance between words, may seem problematic, tests conducted in the *grid search* 4.2 show that incorporating this distance does not significantly improve the results.

4 Optimized Evaluation of Transpositions and Lexical Variations with UTER

To overcome the identified limitations of TER, we propose UTER, an enhanced metric that integrates a reordering algorithm for detecting transpositions, accounts for lexical and morphological variations, and includes a weighting of edit errors.

4.1 Description of the UTER Algorithm

The UTER metric evaluates translation quality based on a series of edit operations and incorporates an enriched lexical approach to better capture morphological nuances.

Like TER, it calculates the number of insertions, deletions, substitutions, and transpositions required. However, it differs by introducing a preliminary reordering phase, where an algorithm optimizes the word order to match the reference, while returning the minimal number of transpositions. This reordering is based on lexical similarity, calculated using the Sørensen-Dice index and controlled by a threshold, allowing the detection and handling of partial lexical matches between words, as illustrated in Figure 2.

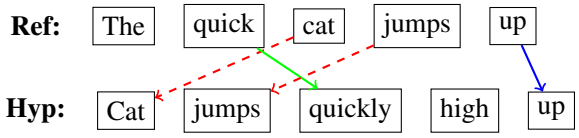

Figure 2: Reordering Algorithm: Blue indicates direct matches, red indicates matches requiring alignment, and green indicates partial matches requiring alignment too. (Minimum transpositions = 2)

After the reordering phase, the edit costs (insertion, deletion, substitution) are calculated using a dynamic programming method similar to the one employed by the classic TER. This approach relies on constructing a matrix where each cell represents the minimal cost to transform a subsequence of the reordered reference into a subsequence of the translation.

At each step, the costs of the three operations (insertion, deletion, substitution) are evaluated, and the corresponding cell is updated with the minimal cost. An adjusted cost is applied to substitutions, calculated based on lexical dissimilarity using the inverted Sørensen-Dice coefficient.

$$\begin{cases} \text{Suppression :} & \text{Cost}[i-1][j]+1, \\ \text{Insertion :} & \text{Cost}[i][j-1]+1, \\ \text{Substitution :} & \text{Cost}[i-1][j-1] \\ & \quad +(1-\text{S}[\text{ref}_{i-1}, \text{hyp}_{j-1}]) \\ \text{where } S \text{ is the Sørensen similarity.} \end{cases}$$

Finally, the backtracking of the matrix illustrated in Figure 3 allows for the precise determination of the required insertions, deletions, and substitutions.

	-	quickly	cat	jumps	up	hight
-	0.0	1.0	2.0	3.0	4.0	5.0
the	1.0	1.0	2.0	3.0	4.0	4.49
quick	2.0	1.10	2	2.53	3.43	4.43
cat	3.0	2.10	1.10	2.10	3.10	4.10
jumps	4.0	3.10	2.10	1.10	2.10	3.10
up	5.0	4.10	3.10	2.10	1.10	2.10

Orange: Match (cost = 0)
Red: Substitution (cost = 0.10)
Green: Deletion (count = 1)
Blue: Insertion (count = 1)

Figure 3: Edit distance matrix and optimal backtracking path between reference and hypothesis in UTER.

The UTER score is then obtained by weighting the total of these edit operations, as well as the transpositions, with coefficients reflecting the relative severity of each type of error. This total is then normalized by the length of the longest sentence, thereby reducing the biases associated with length differences between the translation and the reference.

$$\text{UTER} = \frac{(I \cdot c_i) + (D \cdot c_d) + (S \cdot c_s) + (T \cdot c_t)}{\max(\text{len(reference)}, \text{len(translation)})}$$

- I, D, T: Number of insertions, deletions and transpositions.
- S: Total substitution cost
- C_I, C_D, C_S, C_T: Weights for each type of operation

Table 1: Comparison of Metric Scores for **{The quick cat jumps up}** and **{Cat jumps quickly high up}**

SBERT	GPT	UTER	METEOR	chrF
0.926	0.85	80.0	0.511	46.568

Table 1 shows that UTER produces a score closer to semantic similarity models such as SBERT and GPT compared to traditional metrics like chrF or METEOR.

4.2 Motivation of Algorithmic Choices

To improve translation evaluation compared to the classic TER, UTER adopts algorithmic choices to better capture lexical variations, reduce biases, and fairly weight errors.

- **Sørensen-Dice Similarity:** The Sørensen-Dice measure is favored for its better algorithmic complexity, $O(n + m)$, where n and m represent the lengths of the compared words. Unlike other measures, typically calculated in $O(n \cdot m)$, this lower complexity ensures faster performance and greater efficiency, especially for large corpora.

  Moreover, Sørensen-Dice produces generally lower similarity scores than other measures, with scores close to those of the Character Error Rate (CER). This restrictive behavior is an advantage as it reduces the risk of false matches between words that are only superficially similar. It allows for more reliable detection of relevant lexical alignments, which is particularly useful for handling linguistic variations in multilingual contexts or for morphologically rich languages (Ouzerrout, 2024).

- **Similarity Threshold and Error Weighting:** A similarity threshold is integrated into the transposition calculation phase to ensure that only relevant matches, with a similarity above this threshold, are considered. Furthermore, each type of error (insertion, deletion, substitution, transposition) is weighted by a coefficient reflecting its relative severity in the evaluation process. This methodology aims to more precisely differentiate major errors from minor ones, allowing for a more nuanced evaluation.

  The intuition behind selecting values for the threshold and coefficients suggests, for instance, a minimal threshold of 0.6, as a lower similarity makes strict matching unlikely. Moreover, it seems appropriate to assign higher coefficients to insertions and deletions than to substitutions and transpositions, as the latter are generally perceived as less severe errors. Finally, it is important to note that a substitution, even involving a lexically distinct word, may still preserve some semantic alignment.

To refine this intuition, a parametric exploration was carried out using a *grid search* approach.

This method allowed for exploring different combinations of coefficients and thresholds to identify the configurations yielding the best results in terms of correlation with human annotations and minimizing the mean and median discrepancies. Additionally, a comparison was made between the Sørensen and Jaro similarity functions, with the latter showing higher similarity scores.

The optimal configuration obtained is:

$$\begin{cases} \text{Similarity threshold} = 0.6, \\ \text{Transposition coefficient} = 0.3, \\ \text{Substitution coefficient} = 0.7, \\ \text{Insertion coefficient} = 1.0, \\ \text{Deletion coefficient} = 0.8, \\ \text{Similarity function} = \textit{Sørensen}. \end{cases}$$

These values confirm the initial intuition, validating the proposed choices of thresholds and coefficients. It is also worth noting that, although the substitution coefficient may seem high, it results from considering an adjusted cost for substitutions rather than their simple occurrence.

5 Comparative Evaluation and Results Analysis

To assess the effectiveness of UTER, we used the WMT22 workshop dataset (Koehn et al., 2022), which provides translations, references, and human evaluation scores based on Direct Assessment (DA) for 41 language pairs, with a total of 1.29 million lines.

In order to test UTER in a multilingual context and on morphologically complex languages, we focused our analysis on eight such languages: Finnish, Estonian, Latvian, Lithuanian, Czech, Kazakh, Zulu, and Xhosa, totaling 89,920 lines. Among these, the last three are truly low-resource, adding an additional layer of complexity, while Estonian and Latvian can be considered mid-resource languages with moderate NLP support.

UTER was compared against benchmark metrics such as BLEU, TER, METEOR, and chrF, as well as alternative metrics like CDER (Leusch et al., 2006) and RIBES (Isozaki et al., 2010).

This panel allows, initially, the validation of UTER's relevance by comparing it with other metrics on languages with complex morphological characteristics. The results include, in addition to Pearson and Kendall correlations, the means and medians, in order to better reflect the overall performance of each metric.

Table 2 reports the Pearson and Kendall correlations, as well as mean and median scores, for UTER and other evaluation metrics. While chrF exhibits slightly higher correlation values, UTER achieves scores closest to human annotations in terms of mean and median, indicating more robust behavior overall.

Metric	Pearson	Kendall	Mean	Median
RAW (Human)	-	-	**53.275**	**54.000**
CHRF	**0.430**	**0.291**	47.163	46.623
RIBES	0.250	0.220	61.742	72.074
CDER	0.291	0.193	36.505	35.294
UTER	0.361	0.242	**53.296**	**51.607**
METEOR	0.367	0.245	39.267	38.070
TER	0.299	0.228	20.543	21.429
BLEU	0.265	0.180	10.019	0.000

Table 2: Correlation and score statistics of UTER and baseline metrics against human direct assessment (DA) scores on morphologically complex languages.

This approach is essential, as a high correlation does not necessarily guarantee a robust evaluation, as highlighted by Xiao et al. (2023) (Xiao et al., 2023) and Nimah et al. (2023) (Nimah et al., 2023). Correlation coefficients merely reflect whether two metrics vary in a similar direction, not whether their are close to human judgments.

UTER stands out for its balanced combination of correlation and overall performance, demonstrating its effectiveness for morphologically complex languages.

To better visualize the results on a dataset of nearly 90k examples, the metric curves were plotted 4 with smoothing applied using a window corresponding to 1% of the data.

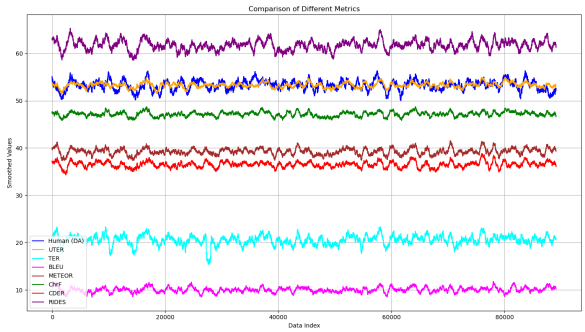

Figure 4: Smoothed curves on 89,920 examples from morphologically complex languages

The UTER metric curve shows a close match with that of the reference human evaluations (Human DA), suggesting a superior ability to reflect human judgments compared to metrics such as BLEU or TER. Furthermore, the chrF metric shows acceptable performance, with notable consistency relative to human evaluations.

In contrast, the RIBES metric appears to tend towards overestimating results, which could limit its reliability.

The Figure 5 provides a closer look at the performance of UTER compared to chrF in approximating human direct assessments.

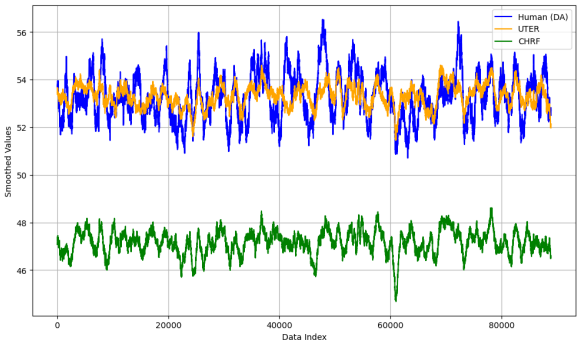

Figure 5: Close-up comparison of UTER and chrF metric curves against human direct assessments on morphologically complex languages.

To assess the relevance of UTER in distinct linguistic contexts, we decided to test its performance on CJK languages (Chinese, Japanese, Korean), which present particular challenges due to their complex writing systems that make classical metrics such as BLEU or TER often unsuitable, due to semantic ambiguities and segmentation variations (Nagata and Morishita, 2020; Zhu, 2020; Song et al., 2020).

This analysis focuses on Chinese and Japanese, totaling 134,115 examples, and concentrates on the metrics that performed best on morphologically complex languages.

Despite being specifically optimized for morphologically complex languages, UTER outperforms other metrics even on languages with complex writing systems such as Chinese and Japanese, as shown in Figure 6, although some areas for improvement remain. In comparison, ChrF produces moderate results, while the CDER and RIBES metrics perform relatively poorly, highlighting their limited relevance in this context.

Figure 6: Comparison of metric performance on 134,115 examples from Chinese and Japanese, illustrating UTER's effectiveness in handling complex writing systems.

Finally, to complete the comparison between UTER and ChrF, we extended the analysis to the entire dataset, totaling 1.29 million lines, with 55% of translations being into English.

The results show 7 that, while the overall performance of UTER and ChrF is moderate on this corpus, UTER maintains better alignment in terms of consistency with human evaluations.

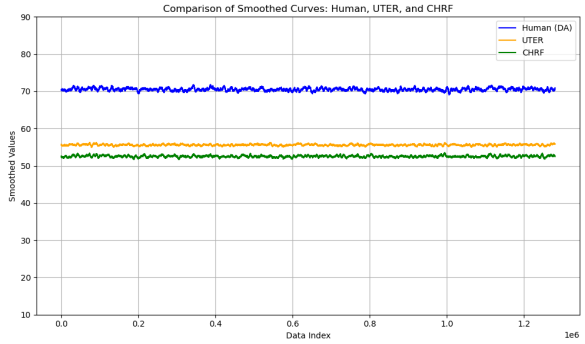

Figure 7: Comparison of UTER and chrF on the full WMT22 dataset (1.29 million lines)

6 Conclusion

In this paper, we presented UTER, a new automatic translation evaluation metric specifically designed to better address the needs of low-resource and morphologically complex languages.

Unlike traditional metrics such as BLEU or METEOR, UTER integrates an optimized reordering algorithm and uses Sørensen-Dice similarity to effectively capture morphological and lexical variations.

Our comparative evaluation, based on a subset of 89,920 lines from the WMT22 dataset, showed that UTER offers a closer match with human evaluations (DA) compared to metrics like chrF, CDER,

or RIBES, particularly for languages with complex linguistic structures.

Expanding the analysis to the entire dataset, which includes 1.29 million lines, with 55% of translations into English, we observed that, although overall performance is moderate on this corpus, UTER continues to outperform ChrF, thus confirming its robustness and ability to adapt to various multilingual contexts.

In conclusion, UTER positions itself as a reliable and versatile metric, offering a fairer evaluation for under-represented languages. Future directions include integrating syntactic information to further refine this metric, particularly for languages with complex grammatical structures.

7 Discussion

Although designed to improve the evaluation of machine translation for under-represented and morphologically complex languages, UTER has certain limitations:

It relies on lexical metrics and an optimized reordering algorithm, but does not account for the deep semantic and contextual aspects that only advanced language models can capture. As a result, UTER may fail to capture certain semantic nuances in translations of long or complex sentences.

Moreover, UTER aims to evaluate the overall performance of translation models by aligning with human evaluations. While it produces results consistent with human judgments in a general evaluation framework, it does not provide a perfect match with human evaluations for every specific case.

The parameters of UTER were specifically optimized for morphologically complex languages, which explains its lower performance on the entire dataset. An improved version of UTER, with parameters adjusted for better performance across a diversity of languages, could be considered.

In summary, UTER provides a reliable estimate of the overall quality of a translation model, but it is not intended to evaluate each individual translation with human-level precision. Future adjustments could strengthen UTER's correlation with human judgments, while aiming for better adaptability in more diverse multilingual contexts.

References

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Philipp Koehn et al. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 241–248, Trento, Italy. Association for Computational Linguistics.

Siegwalt Lindenfelser. 2020. Asymmetrical complexity in languages due to l2 effects: Unserdeutsch and beyond. *Languages*, 5:57.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS ONE*, 5(1):e8559.

Masaaki Nagata and Makoto Morishita. 2020. A test set for discourse translation from japanese to english. In *NTT Communication Science Laboratories*, Kyoto, Japan.

Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.

Samy Ouzerrout. 2024. Universal-WER: Enhancing WER with segmentation and weighted substitution for varied linguistic contexts. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 29–35, Helsinki, Finland. Association for Computational Linguistics.

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

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Anushka Singh, Ananya B. Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M. Khapra. 2024. How good is zero-shot mt evaluation for low resource indian languages? *arXiv preprint arXiv:2406.03893*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Huacheng Song, Yi Li, Yiwen Wu, Yu Liu, Jingxia Lin, and Hongzhi Xu. 2020. How grammatical features impact machine translation: A new test suite for chinese-english mt evaluation. In *Shanghai International Studies University*, Hong Kong, Shanghai.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Zhongyuan Zhu. 2020. Evaluating neural machine translation in english-japanese task. In *Weblio Inc.*