Evaluating Automatic Metrics
with Incremental Machine Translation Systems

Guojun Wu[1]   **Shay B. Cohen**[2]   **Rico Sennrich**[1,2]
[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh
guojun.wu@uzh.ch, scohen@inf.ed.ac.uk, sennrich@cl.uzh.ch

Abstract

We introduce a dataset comprising commercial machine translations, gathered weekly over six years across 12 translation directions. Since human A/B testing is commonly used, we assume commercial systems improve over time, which enables us to evaluate machine translation (MT) metrics based on their preference for more recent translations. Our study not only confirms several prior findings, such as the advantage of neural metrics over non-neural ones, but also explores the debated issue of how MT quality affects metric reliability—an investigation that smaller datasets in previous research could not sufficiently explore. Overall, our research demonstrates the dataset's value as a testbed for metric evaluation. We release our code.[1]

1 Introduction

Automatic metrics for machine translation (MT) are typically assessed by measuring their correlation with or accuracy with respect to human judgments (Macháček and Bojar, 2013; Mathur et al., 2020b; Kocmi et al., 2021). However, human evaluation is resource-intensive and time-consuming, and the number of translation systems included in a meta-evaluation tends to be relatively small. In this study, we explore the use of commercial machine translations, collected weekly over a period of 6 years for 12 translation directions, for the evaluation of MT metrics. Given the common use of human A/B testing (Tang et al., 2010; Caswell and Liang, 2020), our base assumption is that commercial systems show real improvements over time and that we can assess metrics as to whether they prefer more recent MT outputs. Using our dataset, we revisit a number of recent findings in MT metrics research, and find that our dataset supports these.

Trained metrics, developed to directly learn human judgments (Rei et al., 2020; Sellam et al.,

2020), showed notable advancements in correlating with human judgments compared to non-neural metrics like BLEU (Papineni et al., 2002; Freitag et al., 2021). Recent studies (Freitag et al., 2022, 2023) also revealed that neural metrics achieved significantly better correlation with human judgments than non-neural ones and were able to generalize to new domains and challenging test sets. In our experiments, we analyze metric scores over time and evaluate metrics' ability to accurately rank MT systems. Our findings demonstrate that neural metrics show a more consistent upward trend, and achieve higher accuracy than non-neural metrics.

Ma et al. (2019) argued that the correlation between metrics and human judgments significantly decreased when considering only the top-performing systems. To be specific, they assessed the stability of metrics across top-N MT systems, and noticed that the correlation between metric and human scores diminished as N decreased. Mathur et al. (2020a) highlighted the instability of correlations with small N, and instead employed a rolling window of N systems, moving from the worst to the best systems. Due to the limited number of MT systems (typically 10-15 systems) in the datasets, they were unable to confirm that the correlation declines as system quality improves with their approach. Due to the much larger size of our dataset (see Section 3.1), we can achieve more stable results when using the rolling window approach from Mathur et al. (2020a). Our findings reveal that a downward trend is the most common, supporting the results of Ma et al. (2019), although upward or relatively flat trends are also seen in some language pairs.

In WMT23 Metrics shared task (Freitag et al., 2023), human translations received unexpectedly low ratings, which prompted an investigation into using synthetic references as a potential alternative. They found that high-quality synthetic references could produce a stronger correlation between hu-

---

[1]https://github.com/gjwubyron/Evo

man evaluations and metrics compared to human references. In our study, we reexamine the usefulness of synthetic references with three language pairs and find that synthetic references can result in comparable correlation.

2 Related Work

The metrics shared task at WMT (Ma et al., 2018; Mathur et al., 2020b) has played a key role in the development and evaluation of automatic metrics. The annual data collected from WMT's comprehensive human evaluation of the translation task provides an ideal foundation for assessing these metrics. In this event, metrics are ranked based on the correlation calculated by comparing their scores to human ratings.

Machine translation systems that are compared typically come from the same evaluation campaign. An exception to this is the study by Graham et al. (2014), who study longitudinal improvements in machine translation quality from 2007–2012 with human assessments, finding that translation quality of the top submissions to the WMT shared task indeed rose significantly. Our study is similarly longitudinal, but our data stems from a single commercial system, and we do not perform our own human evaluation, but instead assume that improvements over time have been validated by company-internal human A/B testing.

Instead of evaluating metrics through comparison with human judgement, Moghe et al. (2023) explored a complementary approach by correlating metrics with the outcome of downstream tasks. Similarly, our study does not use human judgment directly; instead, we evaluate metrics based on their preference for newer MT outputs.

3 Methods

We turn next to describe our data and the metrics we use.

3.1 Data

The original corpus contains sentences in English from Abstract Meaning Representation (AMR) Annotation Release 2.0 (Knight and et al., 2017), along with their German, Italian, Spanish, and Chinese translations developed by Damonte and Cohen (2018) and released by LDC in Damonte and Cohen (2020). This corpus contains 1371 sentences per language. The source sentences were mainly drawn from content gathered in the news domain.

Translations[2] are gathered weekly from May 2018 – March 2024 using Google Translate from each of the five languages to the other four languages. Early experiments revealed that for English→Spanish, there was a substantial similarity between professional translations and those generated by the earliest systems (details in Appendix A). Consequently, Spanish was removed from further investigation, reducing the number of language pairs to 12. As minimal variation was observed between consecutive weeks, we subsample for the following analysis, with consecutive systems being approximately one month apart. After removing duplicates (systems receiving identical scores across all metrics), we retained 56–63 systems per language pair.

3.2 Metrics

In this section, we outline the three types of metrics used in this study.

3.2.1 Surface-level overlap

BLEU (Papineni et al., 2002) measures n-grams overlap between the translation and its reference. We use *corpus_bleu* in SacreBLEU (Post, 2018). **chrF** (Popović, 2015) assesses the overlap between the characters of the translation and the reference. We use *corpus_chrf* in SacreBLEU.

3.2.2 Embedding based

BERTScore (Zhang et al., 2020) derives contextual embeddings from BERT (Devlin et al., 2019) models and computes cosine similarity between embeddings of the translation and the reference. We use the F1 score without TF-IDF weighting.

3.2.3 Trained with human judgements

COMET-20 (Rei et al., 2020) is trained on top of XLM-R (Conneau et al., 2020) using Direct Assessments (DA) from WMT17 to WMT19. We use wmt20-comet-da. **UniTE** (Wan et al., 2022a,b) is capable of evaluating translation outputs in source-only, reference-only, and source-reference-combined assessment scenarios. We use unite-mup. **COMET-22** (Rei et al., 2022a) is the current default model in COMET and trained on DA from WMT17 to WMT20. We use wmt22-comet-da.

---

[2]Due to the origin of the translations, the data we use is to be licensed by the Linguistic Data Consortium. If you would like to use this dataset for your research, please contact the authors. The collection of the data continues.
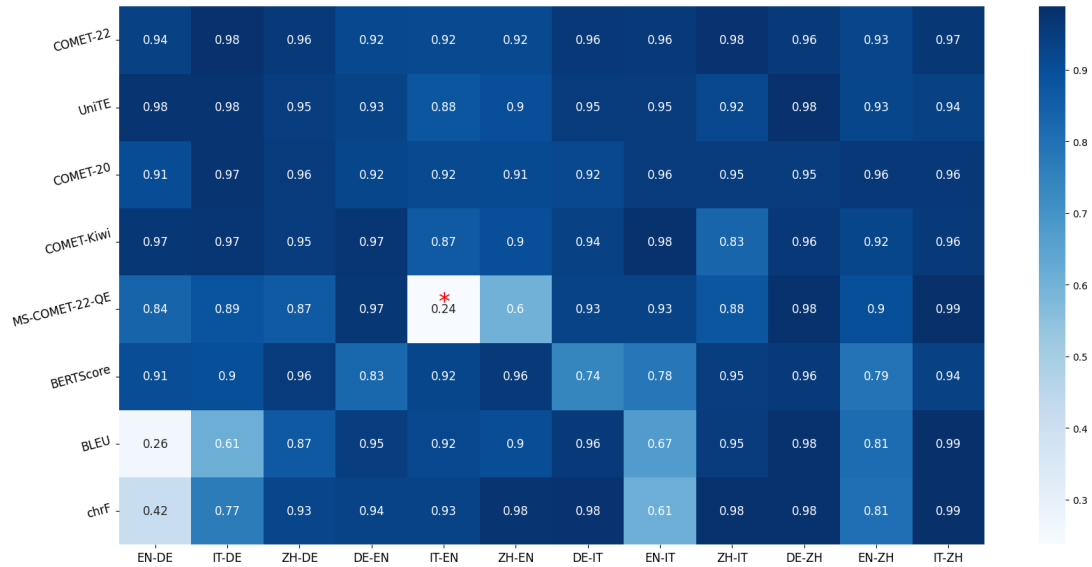
Figure 1: The Spearman correlation measures the relationship between metric score rankings and time rankings for MT systems. A positive correlation indicates an upward trend, with a higher correlation indicating a stronger trend. A red star indicates lack of statistical significance (p-value > 0.05).

COMET-Kiwi (Rei et al., 2022b) is a reference-free metric trained using DA from WMT17 to WMT20, and DA from the MLQE-PE corpus. We use wmt22-cometkiwi-da.

MS-COMET-QE-22 (Kocmi et al., 2022) is a reference-free metric, extending COMET by Microsoft Research with proprietary data.

4 Results

4.1 How do metric scores change over time?

While it is reasonable to expect that systems improve over time, how metric scores will reflect these improvements remains unclear. To investigate this, we visualize how metric scores vary over time for individual language pairs in Appendix B. In general, upward trends are evident for the metrics across the language pairs.

We use Spearman correlation to measure whether the upward trends are consistent. Metrics with higher correlation are deemed more reliable, as they better reflect the overall ranking of the systems. As illustrated in Figure 1, COMET-22, UniTE, COMET-20, and COMET-Kiwi consistently demonstrate high correlation across the language pairs. Among the remaining four metrics, we notice low correlations in specific language pairs, like BLEU and chrF in English→German or MS-COMET-22-QE in Italian→English.

4.2 How good can the metrics rank incremental systems accurately?

In this section, we evaluate metrics in a common scenario (Mathur et al., 2020a): ranking a pair of systems. As we assume newer systems are better than old ones, accuracy (Kocmi et al., 2021) is adopted as follows. For each system pair, we calculate the difference of the metric scores (metricΔ) and the difference in time (timeΔ). Accuracy for a specific metric is calculated as the ratio of rank agreements between metric and time deltas to the total number of comparisons:

$$\text{Accuracy} = \frac{|\text{sign(metric}\Delta) == \text{sign(time}\Delta)\,|}{|\text{all system pairs}|}$$

Since the systems span from 2018 to 2024, those separated by a substantial time interval might exhibit considerable quality gaps, potentially resulting in an overestimate of metric reliability (Mathur et al., 2020a). Consequently, we only pair systems with a gap of less than a year. Even within such a timeframe, substantial improvements in quality are possible (Caswell and Liang, 2020).

Table 1 shows that trained metrics generally outperform non-trained metrics. For all system pairs, COMET-22 achieves the highest accuracy, followed by COMET-Kiwi. In contrast, MS-COMET-QE-22 struggles to attain high accuracy except for into Chinese. Among surface-level metrics, chrF outperforms BLEU, reflecting results in previous

	All	Into EN	From EN	Into DE	Into IT	Into ZH
COMET-22	**73.9**	66.6	71.6	76.4	**79.4**	72.6
COMET-Kiwi	73.4	72.1	73.9	74.8	75.3	71.4
UniTE	73.2	66.5	73.7	**77.1**	75.0	73.9
COMET-20	72.5	66.1	**74.6**	74.3	74.0	74.9
chrF	71.4	**74.5**	57.8	60.4	76.5	74.6
MS-COMET-22-QE	69.9	57.4	68.1	68.8	73.9	**78.6**
BLEU	68.2	71.7	57.3	56.3	68.9	76.4
BERTScore	68.0	65.4	62.2	68.8	69.0	68.6

Table 1: Accuracy for ranking system pairs. Column "All" shows the results for all system pairs. Each following column evaluates accuracy over a subset of systems. Rows are sorted by the accuracy over all system pairs.
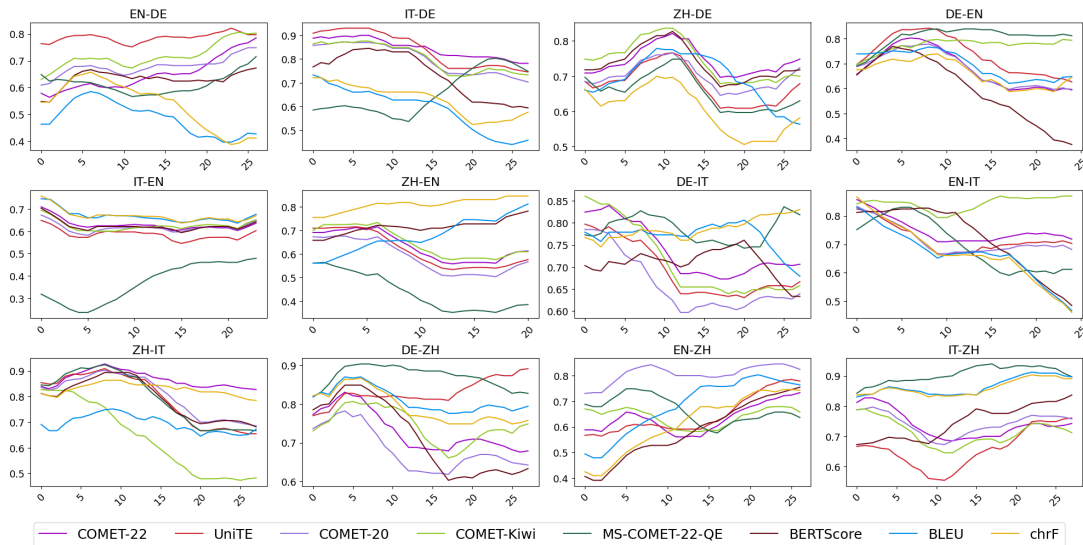

Figure 2: Accuracy over a rolling window of 36 systems. The x-axis represents the index of the starting system, with systems ordered chronologically from the earliest to the most recent. The x-axis scale may vary due to differing numbers of systems, as discussed in Section 3.1.

studies (Kocmi et al., 2021), and achieves the highest accuracy for into English. We also examine performance for individual language pairs. Trained metrics exhibit high accuracy, yet no single metric excels across all pairs. More details in Appendix C.

4.3 Does the reliability of metrics depend on the quality of the systems evaluated?

As mentioned in Section 1, metric reliability may decline as the quality of evaluated systems improves (Ma et al., 2019). However, the limited number of MT systems made it difficult to fully confirm this trend (Mathur et al., 2020a). We revisit this issue using a larger sample of MT systems. Following the approach of Mathur et al. (2020a), we implement a rolling window of N systems, transitioning from the earliest to the most recent ones.

Using accuracy as explained in Section 4.2, we conduct tests with N varying from 24 to 40. Fig-

ure 2 illustrates the results for N = 36, representing the identified scenarios. Different metrics display varying trends. For instance, in English→German, trained metrics show an upward trend, while surface-level metrics show a downward trend. A downward trend is most common, with each metric showing a clear decline across 7 or more language pairs. However, we also observe upward or relatively flat trends in the remaining language pairs.

4.4 Can synthetic references serve as an alternative to human references?

We generate synthetic references for three language pairs using another commercial MT system, DeepL, and examine their impact on metric evaluation. As depicted in Figure 3, we observe that for English→German, all metrics achieve a higher accuracy, while for the remaining language pairs,
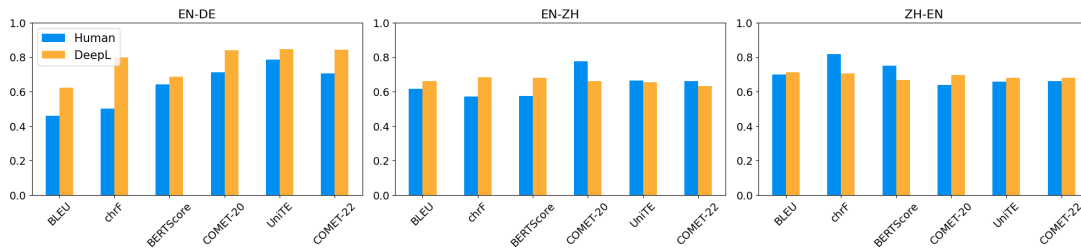
Figure 3: Accuracy across three language pairs using either human or synthetic references. The two reference-free metrics are not included as they will not be influenced by reference.

there are some drops. Overall, synthetic references lead to a comparable accuracy for the three language pairs we investigate, suggesting that they can be used when human references are unavailable.

5 Conclusion

Our dataset, covering 12 language pairs with at least 56 machine translations each, surpasses previous datasets that typically included only 3 pairs with around 15 machine translations each. Based on the assumption that newer translations of a regularly updated commercial system tend to be of a higher quality, we apply the dataset to revisit prior findings on MT metrics. We provide larger-scale evidence for debated questions such as the relationship between MT quality and metric reliability—issues that previous research was unable to conclusively resolve on smaller datasets. Additionally, the systems are incremental (a baseline compared to improvements developed by the same group), reflecting the most common use case of the metrics. We encourage the use of our dataset for future investigations into MT metrics or the development of MT quality over time.

Limitations

Our study bases on the assumption that newer systems of Google Translate outperform older ones due to the quality assurance measures, including human testing, taken before deployment. Although this is a reasonable belief, it might not always be true.

Recently, LLM-based evaluators have demonstrated great performance in evaluating MT systems. However, we have not included any LLM-based evaluators in this study because it would be costly to experiment with our extensive dataset.

References

Isaac Caswell and Bowen Liang. 2020. Recent advances in google translate. https://research.google/blog/recent-advances-in-google-translate/. Google Research Blog.

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

Marco Damonte and Shay Cohen. 2020. Abstract Meaning Representation 2.0 - Four Translations LDC2020T07. Web Download. Philadelphia: Linguistic Data Consortium, 2020.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Kevin Knight and et al. 2017. Abstract Meaning Representation (AMR) Annotation Release 2.0 LDC2017T10. Web Download. Philadelphia: Linguistic Data Consortium, 2017.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, 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. Association for Computational Linguistics.

Ricardo Rei, 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. 2022b. 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). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings 16th Conference on Knowledge Discovery and Data Mining*, pages 17–26, Washington, DC.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022a. Alibaba-translate China's submission for WMT2022 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 586–592, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022b. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Appendices

A Metric scores for English → Spanish translations

Figure 4 displays the scores of four different metrics for English→Spanish translations in our early experiments. Early systems achieved nearly perfect metric scores, whereas later systems displayed markedly lower scores. Upon closer examination of the human translations, we noticed roughly 25% of them are identical to that of the early systems. This indicates the use of Google Translate in the professional translation.
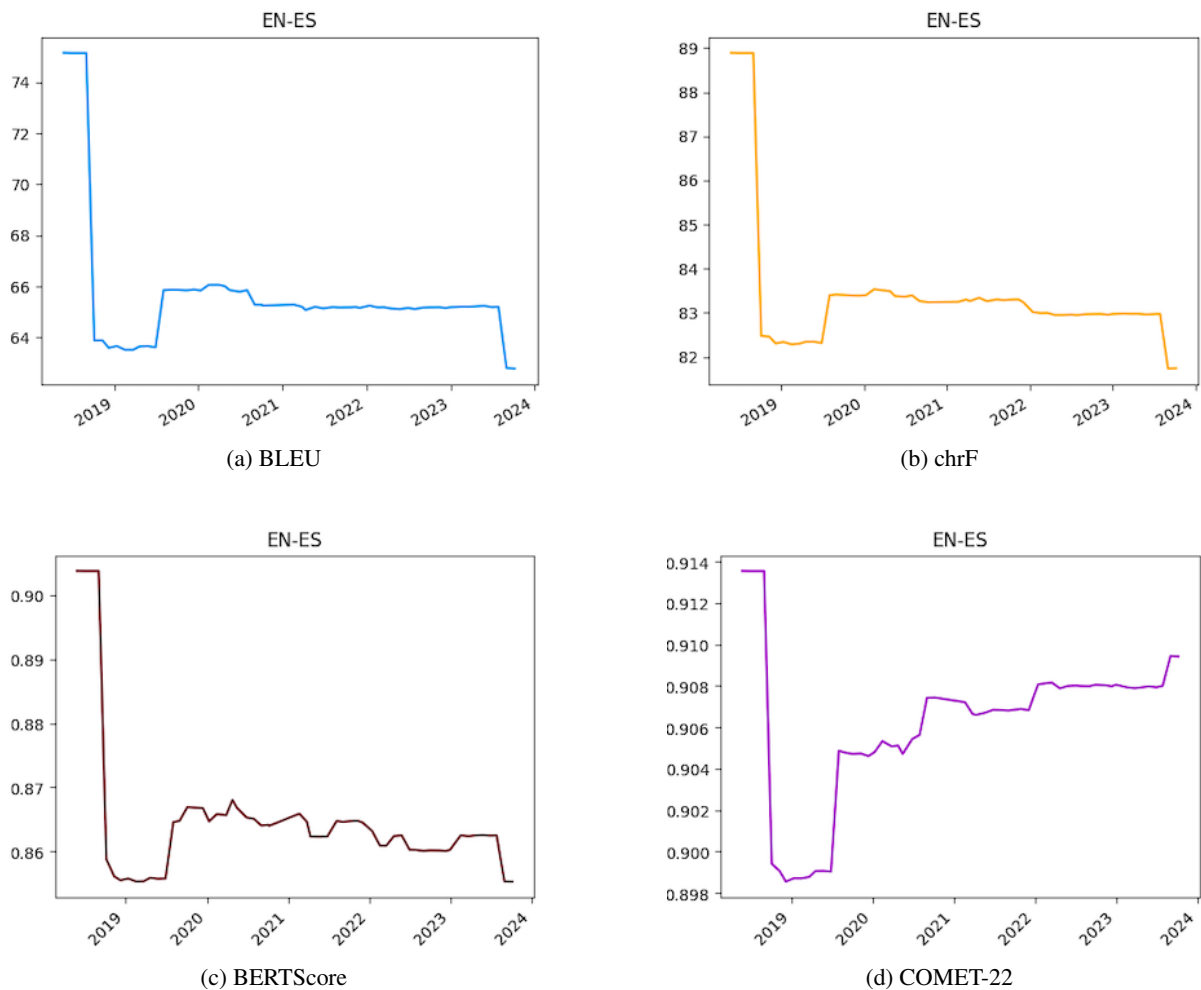

(a) BLEU

(b) chrF

(c) BERTScore

(d) COMET-22

Figure 4: The metric scores for English→Spanish translations. While the earliest system achieved nearly perfect scores, subsequent systems showed a notable decline.

B Metric scores over time

Figure 5 illustrates the findings regarding the change of metric scores over time. Generally, upward trends are evident for the metrics across language pairs. Furthermore, these trends sometimes appear as step-like progressions. Based on a visual inspection of the results, we have some interesting findings as follows:

1. Although there have been concerns that MT systems were optimized for BLEU, given its longstanding status as the primary evaluation metric, our findings suggest that the upward trends of BLEU are less consistent compared to other metrics. This observation might provide implicit evidence that BLEU is not solely used during system development.

2. The trajectories of BLEU and chrF exhibit a high degree of similarity, as do the trajectories of COMET-20, COMET-22, COMET-Kiwi, and UniTE. In contrast, BERTScore and MS-COMET-22-QE follow distinct trajectories of their own. These similarities and discrepancies reflect the inherent properties of these metrics. BLEU and chrF both rely on measuring surface-level overlap, while BERTScore is unique in its reliance on contextual embeddings. As for the trained metrics, although they are all trained in a similar manner, MS-COMET-22-QE was trained using entirely different data.

3. In certain language pairs, the trajectories of certain metrics may experience a downturn. For instance, noticeable troughs are observed for BLEU and chrF in English→German, Italian→German, and English→Italian; for BERTScore in English→German, German→Italian, and English→Italian; and for MS-COMET-22-QE in Italian→English, Italian→German, and Chinese→English. On the other hand, the trajectories of the remaining metrics may occasionally exhibit bumps but do not show clear troughs.
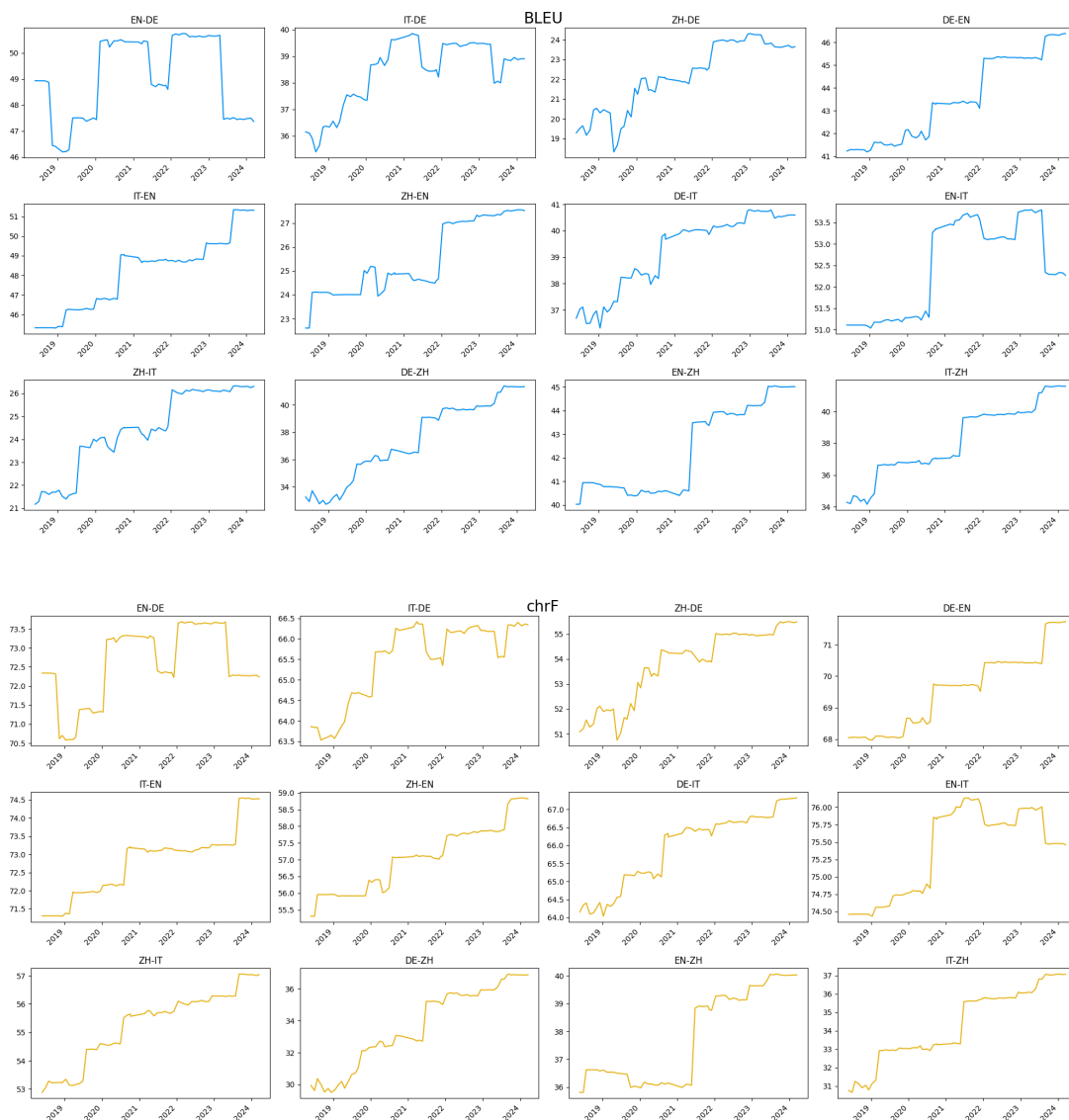

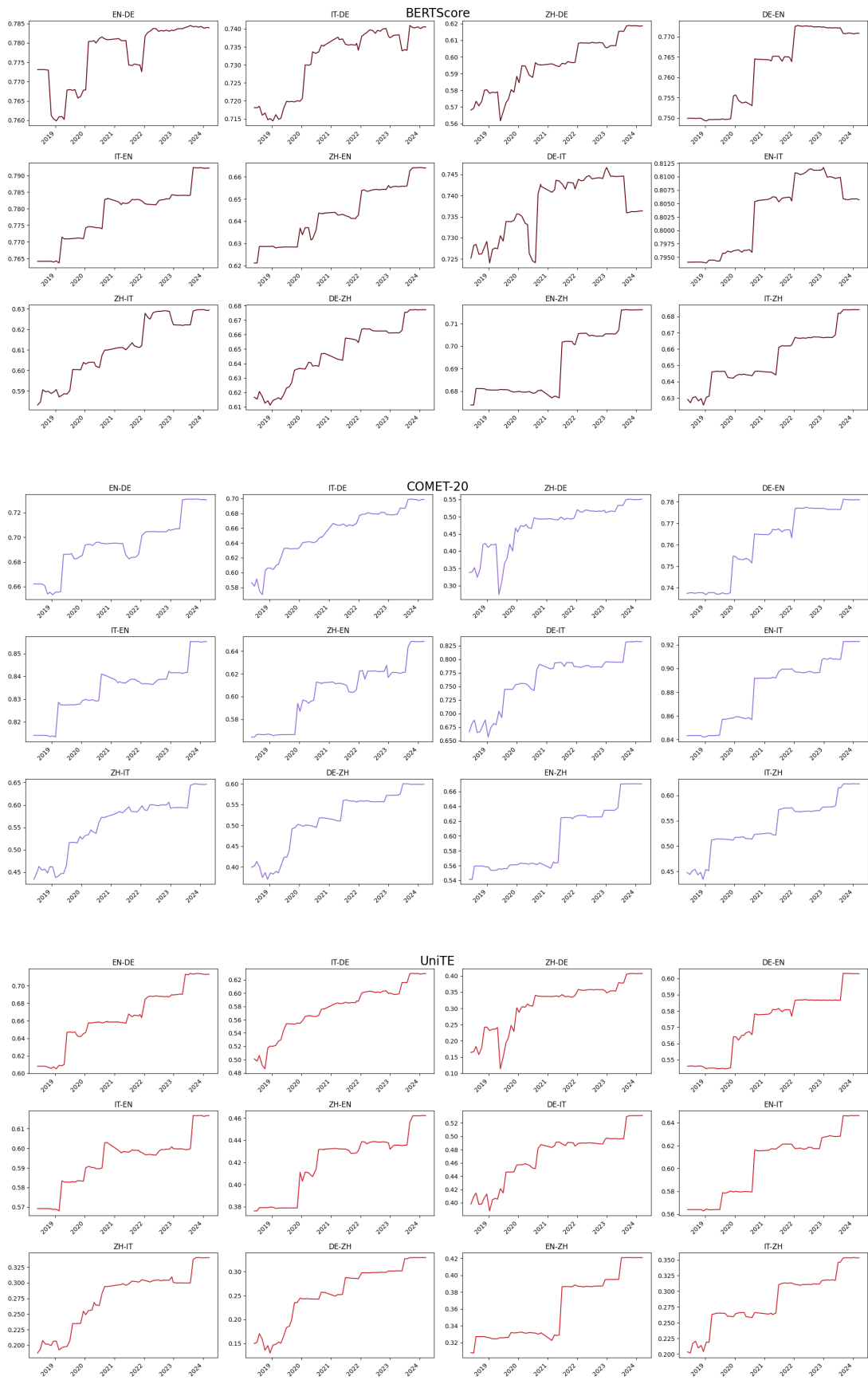Figure 5: Metric scores over time.
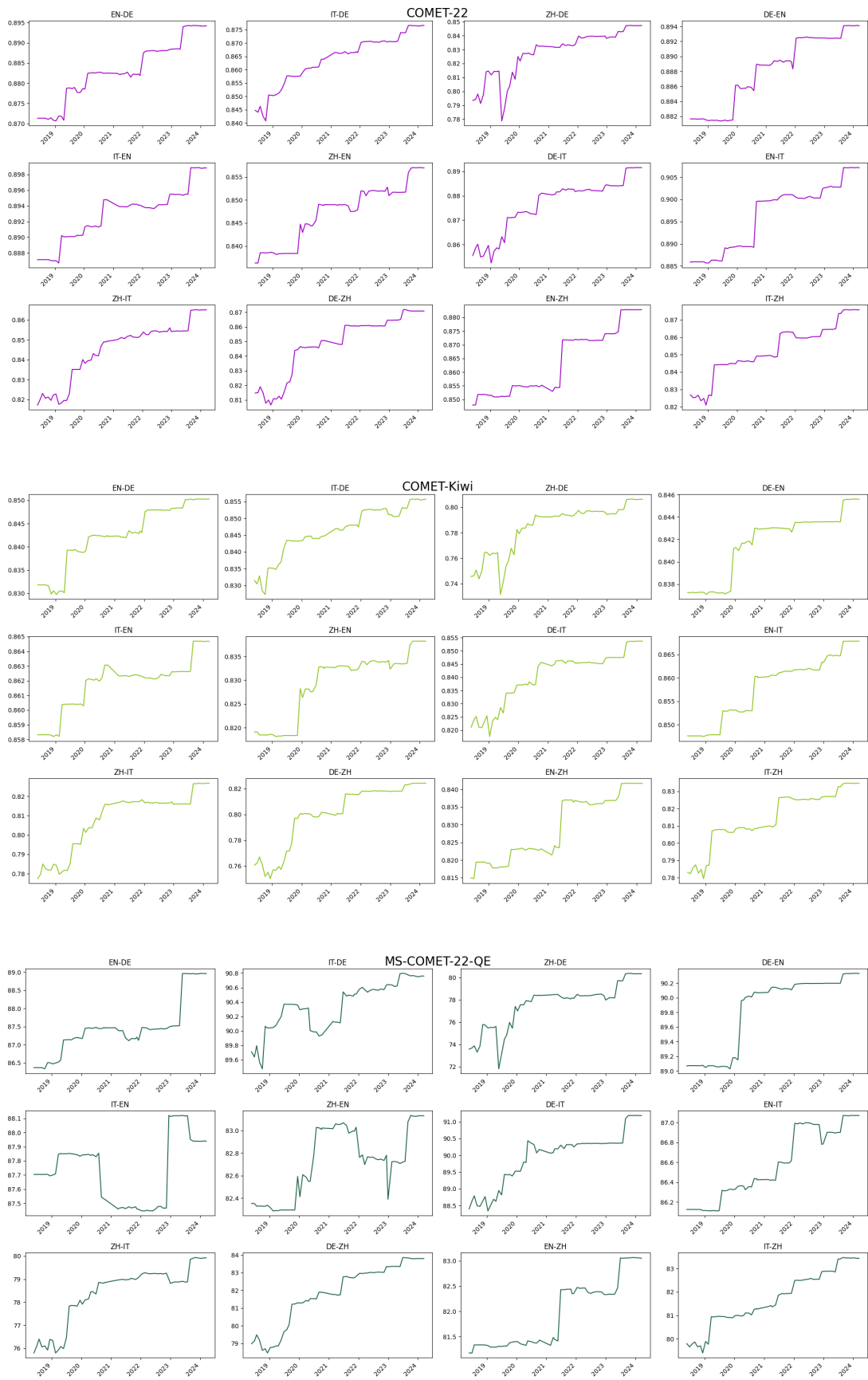
Figure 5: Metric scores over time.

Figure 5: Metric scores over time.
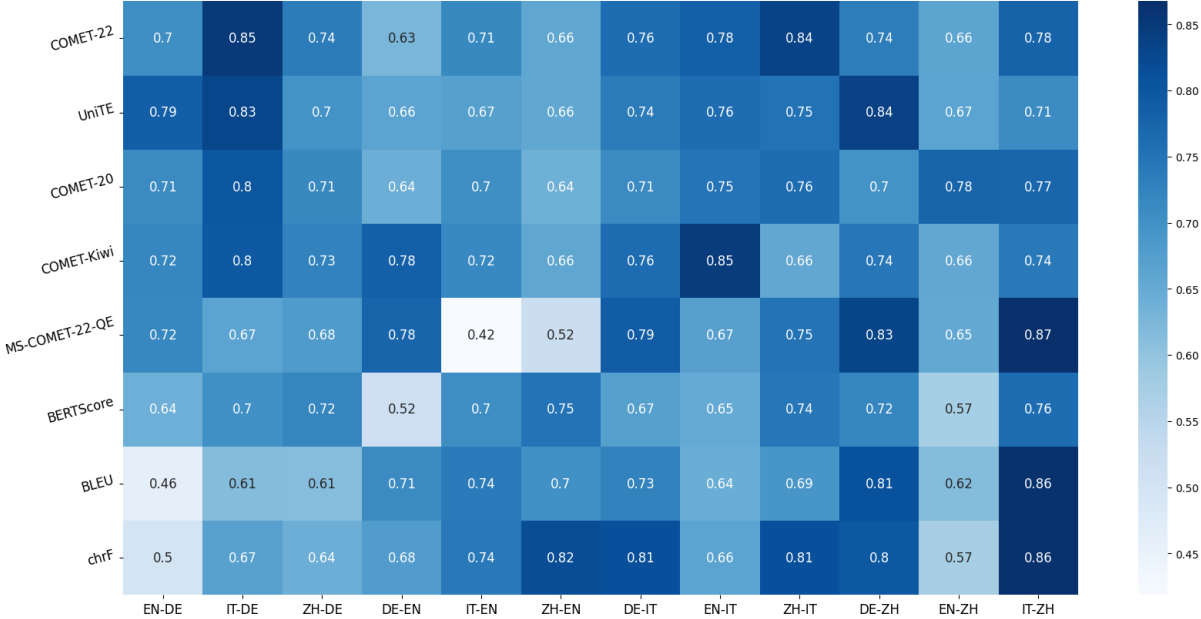
C Accuracy across language pairs

Figure 6: Accuracy for ranking system pairs across individual language pairs.